PARMY OLSON

SUPREMACY

L'AI, ChatGPT e la sfida che cambierà il mondo

Garzanti

PARMY OLSON

SUPREMACY

L'AI, ChatGPT e la sfida che cambierà il mondo

Garzanti

L’autrice

Parmy Olson, oggi editorialista di «Bloomberg», ha scritto per «Forbes» e «The Wall Street Journal» e da anni si occupa di intelligenza artificiale, social media e regolamentazione tecnologica. Nel 2019 «Business Insider» l’ha inserita tra le 100 personalità più importanti del Regno Unito nel settore tecnologico, e nel 2023 è stata nominata giornalista digitale dell’anno dalla Public Relations Consultants Association. *Supremacy* è stato nominato libro vincitore del Financial Times and Schroders Business Book of the Year Award 2024.

Parmy Olson

SUPREMACY

L’AI, ChatGPT e la sfida che cambierà il mondo

Traduzione di Giuseppe Maugeri

Garzanti

www.garzanti.it

IL LIBRAIO
www.illibraio.it

Progetto grafico di copertina: *the*World*of*DOT

Traduzione dall'inglese di
Giuseppe Maugeri

Titolo originale dell'opera:
Supremacy

© 2024 by Parmy Olson
Published by arrangement with St. Martin's Publishing Group
All rights reserved

ISBN 978-88-11-01949-7

© 2025, Garzanti S.r.l., Milano
Gruppo editoriale Mauri Spagnol

Prima edizione digitale: giugno 2025

Quest'opera è protetta dalla Legge sul diritto d'autore.
È vietata ogni duplicazione, anche parziale, non autorizzata.
È vietato inoltre l'utilizzo della presente opera per attività di estrazione dati, apprendimento automatico, e qualsiasi altra forma di analisi automatizzata dei dati senza il previo consenso scritto dell'editore, detentore dei diritti di pubblicazione. L'utilizzo non autorizzato per tali scopi costituisce una violazione dei diritti d'autore e sarà perseguito in conformità alle leggi vigenti in materia.

Indice

Prologo

ATTO I - IL SOGNO
- 1. Un eroe tra i banchi di scuola
- 2. Vincere, vincere, vincere
- 3. Salvare gli esseri umani
- 4. Un cervello migliore
- 5. Per l’utopia, per i soldi
- 6. La missione

ATTO II - I LEVIATANI
- 7. Giochi e giochetti
- 8. È meraviglioso
- 9. Il paradosso di Golia

ATTO III - I CONTI
- 10. Le dimensioni contano
- 11. Vincolati a Big Tech
- 12. Sfatare i miti

ATTO IV - LA CORSA
- 13. Ciao, ChatGPT
- 14. Un vago presentimento di sventura
- 15. Scacco matto
- 16. All’ombra dei monopoli

Ringraziamenti

Fonti

SUPREMACY

A Mani

PROLOGO

Dopo aver letto le prime parole di questo libro, potreste chiedervi se siano state scritte da un essere umano.

Nessun problema. Non mi offendo.

Due anni fa, un dubbio del genere non vi avrebbe nemmeno sfiorato. Ma oggi le macchine generano articoli, libri, illustrazioni e codici informatici indistinguibili dai contenuti creati dagli umani. Avete presente la «macchina per scrivere romanzi» del futuro distopico di *1984* di Orwell o, per rimanere nella stessa opera, il «versificatore» che componeva musica popolare? Be', oggi esistono davvero, e il cambiamento che li ha prodotti è avvenuto così in fretta da lasciarci tutti storditi e in preda al timore che, nel giro di un anno o due, parecchia gente perderà il posto di lavoro. Milioni di professionisti del settore impiegatizio sembrano improvvisamente vulnerabili. Giovani illustratori di talento si chiedono se valga ancora la pena iscriversi a una scuola d'arte.

L'aspetto più straordinario è appunto la rapidità con cui è avvenuto tutto questo. Nei quindici anni in cui ho scritto sull'industria tecnologica, non ho mai visto un settore evolversi altrettanto velocemente dell'intelligenza artificiale negli ultimi due anni. Il lancio di ChatGPT, nel novembre del 2022, ha innescato una corsa alla creazione di un nuovo tipo di AI, capace non solo di processare informazioni, ma anche di *generarne*. All'epoca, gli strumenti di AI potevano produrre immagini sbilenche di cani. Adesso generano ritratti iperrealistici di Donald Trump, con imperfezioni e pori della pelle così credibili da rendere quasi impossibile distinguere il falso dal vero.

Molti sviluppatori di AI sostengono che questa tecnologia possa aprire la strada a una sorta di utopia. Altri, invece, temono che porterà al collasso della civiltà per come la conosciamo. In realtà, gli scenari fantascientifici ci hanno distolto dalle minacce più subdole dell'AI (ovvero dal modo in cui alimenta il razzismo, per esempio, o in cui pone a rischio interi settori creativi).

Dietro a questa forza invisibile ci sono aziende che hanno preso il controllo dello sviluppo dell'AI e si sono lanciate in una corsa sfrenata per renderla sempre più potente. Spinte da un'insaziabile voglia di crescere, queste aziende hanno aggirato regole e ingannato il pubblico riguardo ai loro prodotti, assumendo – in maniera decisamente discutibile – il ruolo di custodi dell'AI.

Nessuna organizzazione nella storia ha mai accumulato tanto potere o raggiunto

un numero così elevato di persone come i colossi tecnologici di oggi. Google gestisce il 90 per cento delle ricerche sul web a livello globale, mentre il software di Microsoft è usato dal 70 per cento dei possessori di un computer. Eppure, nessuna delle due aziende si accontenta. Microsoft vuole accaparrarsi una fetta di quel mercato delle ricerche online (un settore da 150 miliardi di dollari) dominato da Google. Google, a sua volta, punta al business del cloud, che frutta a Microsoft 110 miliardi di dollari. In questa battaglia, entrambe hanno attinto alle idee della controparte, al punto che, riducendo il tutto ai minimi termini, il nostro futuro con l'AI è stato scritto, di fatto, da soli due uomini: Sam Altman e Demis Hassabis.

Il primo, esile quanto imperturbabile, è un imprenditore sulla soglia dei quarant'anni che va in ufficio con le sneakers. L'altro è un ex campione di scacchi, ormai prossimo ai cinquant'anni, con una fissazione per i giochi. Entrambi sono leader di grande intelligenza e carisma, capaci di delineare visioni di un'AI onnipotente così affascinanti da attirare una devozione quasi settaria. Entrambi sono arrivati al vertice perché mossi da un'unica ossessione: vincere. Altman è il motivo per cui il mondo ha avuto ChatGPT. Hassabis è il motivo per cui lo ha ottenuto così in fretta. Il loro percorso non solo ha definito la corsa attuale all'AI, ma ha anche plasmato le sfide che ci attendono, incluso l'arduo compito di garantire un futuro etico all'AI in un settore dominato dai giganti dell'industria.

Hassabis si è esposto al rischio di essere deriso dalla comunità scientifica quando ha fondato DeepMind, la prima azienda al mondo con l'obiettivo dichiarato di creare un'AI paragonabile all'intelligenza umana. Il suo intento era fare scoperte rivoluzionarie sulle origini della vita, sulla natura della realtà e sulle cure per le malattie. «Risolvi il problema dell'intelligenza», diceva, «e risolverai tutto il resto.»

Qualche anno dopo, Altman ha fondato OpenAI con la stessa ambizione, ma con un'attenzione maggiore all'obiettivo di portare abbondanza economica all'umanità, far crescere il benessere materiale e migliorare la qualità della vita. «Questo potrebbe essere lo strumento più prezioso mai creato dall'uomo», mi ha detto. «Permetterà a ciascuno di noi di fare cose ben al di là di ciò che oggi ci sembra possibile.»

I loro progetti superavano persino quelli dei visionari più audaci della Silicon Valley. Puntavano a creare un'AI talmente potente da trasformare la società e rendere obsolete intere discipline, dall'economia alla finanza. E solo a loro due, Altman e Hassabis, sarebbe toccato il compito di elargire i frutti di questa rivoluzione.

Nel loro tentativo di realizzare quella che potrebbe essere l'ultima invenzione dell'umanità, entrambi si sono trovati ad affrontare il problema di come controllare una tecnologia così trasformativa. Inizialmente, ritenevano che colossi tecnologici come Google e Microsoft non dovessero guidarne lo sviluppo, poiché troppo orientati al profitto a scapito del benessere collettivo. Così, per anni, sulle due sponde dell'Atlantico, entrambi hanno proceduto a tentoni nella speranza di trovare nuove soluzioni per strutturare i laboratori di ricerca con lo scopo di proteggere l'AI e garantirne un utilizzo orientato al bene. La promessa, dunque, era di agire come

custodi responsabili.

Ma ambedue volevano bruciare l'altro sul tempo. Per costruire il software più potente della storia, servivano denaro e potenza di calcolo, e la risorsa migliore rimaneva la Silicon Valley. Con gli anni, sia Altman sia Hassabis si sono convinti di avere bisogno dei giganti della tecnologia. I nuovi traguardi nello sviluppo di un'AI superintelligente e l'emergere di nuove ideologie in contrasto con la loro visione li hanno portati a un compromesso rispetto ai nobili propositi originari. Entrambi hanno dunque ceduto il controllo a grandi aziende che si sono precipitate a vendere strumenti di AI al pubblico, con una scarsissima supervisione da parte dei regolatori che ha generato conseguenze di vasta portata.

La concentrazione di potere nell'AI avrebbe ridotto la concorrenza e inaugurato nuove ingerenze nella vita privata, oltre che nuove forme di pregiudizio razziale e di genere. Oggi, se chiedete a un popolare strumento di AI di generare immagini di donne, questo vi restituirà una raffigurazione sexy in abiti succinti; alla richiesta di immagini fotorealistiche di amministratori delegati, produrrà ritratti di uomini bianchi; se invece gli domandate di mostrarvi un criminale, spesso genererà immagini di uomini neri. Ciononostante, questi strumenti si stanno insinuando nei nostri feed social, negli smartphone e perfino nei sistemi giudiziari, senza la dovuta attenzione a come potrebbero influenzare l'opinione pubblica.

Il percorso di Altman e Hassabis non è stato molto dissimile da quello seguito un paio di secoli fa da altri due imprenditori, Thomas Edison e George Westinghouse, in guerra l'uno contro l'altro. Entrambi avevano perseguito il sogno di creare un sistema dominante che distribuisse l'elettricità a milioni di consumatori. Entrambi erano inventori diventati imprenditori e sapevano che la loro tecnologia, un giorno, avrebbe alimentato il mondo moderno. La domanda era: quale versione della tecnologia avrebbe prevalso? Alla fine, lo standard elettrico più efficiente di Westinghouse ottenne maggiore diffusione a livello globale. Ma non fu lui a vincere la cosiddetta «guerra delle correnti», perché a imporsi fu General Electric.

In un quadro in cui gli interessi aziendali spingevano Altman e Hassabis a lanciare modelli sempre più onnicomprensivi e potenti, sono stati i giganti tecnologici a emergere come veri vincitori; solo che, questa volta, la corsa puntava a replicare la nostra stessa intelligenza. Oggi il mondo è in balia di un cambiamento epocale. L'AI generativa promette di rendere le persone più produttive garantendo l'accesso a informazioni rilevanti tramite strumenti come ChatGPT. Ma ogni innovazione ha un prezzo. Aziende e governi si stanno adattando a una nuova realtà in cui distinguere tra ciò che è reale e ciò che invece è «generato dall'AI» rappresenta un vero salto nel buio. Le aziende investono somme spropositate in software di AI per sostituire i propri dipendenti e aumentare i margini di profitto. E sta emergendo una nuova generazione di dispositivi personali di AI in grado di condurre un livello di sorveglianza personale mai immaginato prima.

La seconda parte del libro esporrà questi rischi, ma prima intendo spiegare come

siamo arrivati a questo punto e come le visioni di due innovatori che speravano di creare un'AI per il bene comune siano state infine travolte dalle forze del monopolio. La loro è una storia di idealismo (ma anche di ingenuità e istanze dettate dall'ego) e racconta una volta di più come sia praticamente impossibile mantenere un codice etico nelle bolle di Big Tech e Silicon Valley. Altman e Hassabis si sono arrovellati a lungo sulla gestione dell'AI, consapevoli di come tale tecnologia andasse maneggiata in modo responsabile per evitare danni irreversibili. Ma non potevano forgiare un'AI con poteri quasi divini senza le risorse delle più grandi aziende tecnologiche del pianeta. Nel tentativo di migliorare la vita umana, hanno finito per rafforzare proprio queste aziende, lasciando il benessere e il futuro dell'umanità al centro di una contesa per la supremazia aziendale. Ed ecco come sono andate le cose.

ATTO I

IL SOGNO

1. UN EROE TRA I BANCHI DI SCUOLA

Sam Altman sapeva che avrebbe dovuto tenere la bocca chiusa. Nella roccaforte conservatrice di St. Louis, in Missouri, la gente non parlava del proprio orientamento sessuale. Mentre il resto dell'America iniziava a confrontarsi con i diritti delle persone gay, nei primi anni Duemila la sua città natale, nel Midwest, era ancora in ritardo sul tema e considerava un crimine andare a letto con qualcuno del proprio sesso. Gli adolescenti come lui, che cominciavano a intuire di essere omosessuali, tendevano a rifugiarsi nel silenzio. Altman era diverso. Doveva parlarne, non già perché voleva che si conoscesse ogni dettaglio della sua vita, ma perché la sentiva come una missione.

Alle superiori, Altman era il genere di ragazzo che sembrava magicamente trascendere le etichette che gli altri cercavano di affibbiargli. Era brillante quanto un qualsiasi secchione, carismatico quanto un qualsiasi atleta. Nei compiti di letteratura inglese emulava la prosa ardua di Faulkner, per poi destreggiarsi a meraviglia nel calcolo matematico. Quindi si tuffava in piscina per impartire ordini alla squadra di pallanuoto, di cui era capitano, o tornava a casa per coordinare lunghe sessioni di videogame con gli amici. A tavola, con i fratelli minori, Max e Jack, fantasticava a lungo di viaggi spaziali e razzi, e quando giocavano a giochi da tavolo, come *Samurai*, si autoproclamava leader. In queste e molte altre situazioni, insomma, gli piaceva stare al comando.

Altman era cresciuto in una famiglia ebrea della classe media: la madre, Connie, faceva la dermatologa, e il padre, Jerry, era un avvocato. Jerry stava contribuendo a promuovere l'edilizia popolare a St. Louis, così come la ricostruzione di edifici storici, e le sue gesta alimentavano nel figlio una visione orientata al bene comune. Sam ricorda ancora il giorno in cui il padre lo portò nel suo ufficio dicendogli qualcosa del tipo: «Bisogna sempre trovare il tempo per aiutare gli altri, a prescindere dai propri impegni».

Primo di quattro fratelli, Sam era molto sicuro di sé e ostentava una spavalderia che suscitava un certo rispetto. Parlava apertamente della propria sessualità quando altri ragazzi della sua età – siamo alla fine degli anni Novanta – la tenevano segreta. Abbracciando un aspetto che molti, nel Midwest, consideravano negativo, finiva in qualche modo per renderlo interessante, in parte con l'intento di aiutare altre persone che si trovavano nella sua stessa situazione.

La chiamata arrivò da internet. Nel momento in cui iniziò a connettersi al portale

web America Online (AOL), Altman si rese conto di quante altre persone vivevano nella sua condizione. Stava proprio in questo la magia di collegarsi alle chat room di AOL: sentivi il tono del dial-up e i *bip* che segnalavano l'accesso mentre il modem stabiliva una connessione sicura con il World Wide Web, seguiti dallo stridore metallico che sembrava provenire da una radio CB fuori uso. E finalmente eri connesso, con il cuore che accelerava i battiti per le possibilità che tutte quelle chat room ti schiudevano: chattare con un'altra persona stimolante seduta davanti a un computer in chissà quale altra parte del mondo! Le chat room avevano nomi come «Beach Party» o «The Breakfast Club». Le più affollate erano confusionarie e piene di strani personaggi, ma se esploravi categorie più specifiche come Amanti degli animali, Fan di *X-Files* o Gay & Lesbo, trovavi conversazioni più pertinenti.

Per persone come Altman, queste chat room divennero una vera e propria ancora di salvezza. Potevi proteggerti grazie all'anonimato di un nickname e seguire conversazioni in cui altri suggerivano luoghi di ritrovo LGBTQ-friendly. Offrivano un senso di appartenenza in un mondo in cui era facile sentirsi emarginati. «Le chat room su AOL sono state trasformative», avrebbe ammesso più tardi alla rivista «New Yorker». «È brutto avere segreti quando hai undici o dodici anni.»

Le chat room di AOL erano così significative per la comunità LGBTQ che nel 1999, quando Altman aveva quattordici anni, circa un terzo di queste erano dedicate a tematiche gay. Al momento del coming out con i genitori, a sedici anni, la madre rimase scioccata. Suo figlio le era sempre sembrato «unisessuale e tecnologico», avrebbe confessato sempre al «New Yorker», anche se sfuggiva alle classificazioni. Tanto per fare un esempio, era vegetariano in una zona dell'America in cui tutti adoravano il barbecue. Era ossessionato dai computer, senza tuttavia essere introverso o socialmente inadeguato. E mentre tutti gli altri ascoltavano musica pop anni Novanta, lui preferiva quella classica.

Gli Altman trasferirono il loro precoce adolescente alla John Burroughs, una scuola privata d'élite nella periferia di St. Louis, famosa per l'ampio campus immerso nel verde e per l'impegno a sviluppare il talento degli studenti «con l'obiettivo di migliorare la società».

Altman si sobbarcò quante più posizioni di leadership gli fu possibile. Oltre che capitano della squadra di pallanuoto, era anche responsabile dell'annuario scolastico e oratore alle assemblee. Socializzava con i docenti e, ogni tanto, si divertiva a infrangere le regole per suscitare un po' di clamore. Una volta, durante il consueto *pep rally* autunnale, Altman e la squadra di pallanuoto si spogliarono sul palco per rimanere in costume e sorridere alle acclamazioni e agli applausi.

Finito nei guai per questo con il direttore sportivo dell'istituto, invece di lasciar perdere e lamentarsi dell'ammonizione con gli amici, o magari con un altro docente, Altman decise di reagire scomodando direttamente i piani alti e bussando alla porta del preside, Andy Abbott, un ex insegnante di inglese dall'aria mite. L'anziano educatore era affascinato da quell'adolescente alto e magro, con i capelli scuri e gli

occhi grandi, che non di rado si presentava nel suo ufficio per proporre idee o lamentarsi di un'ingiustizia che intendeva trattare sul giornale della scuola.

Il giovane non si lasciava intimidire dall'autorità. Se il preside prendeva una decisione impopolare riguardo ad altri studenti, il ragazzo si sentiva in dovere di soccorrerli. «Si opponeva», racconta Abbott, «e lo faceva in modo ponderato.» A oggi, il pacato educatore considera Altman come «il ragazzo più brillante mai conosciuto».

Questa stessa sincerità, insieme alla capacità di mostrarsi diretto e vulnerabile, avrebbe in seguito aiutato il giovane a conquistarsi la fiducia di altre figure influenti nel mondo della tecnologia e della politica, o presso gli investitori e la stampa, per arrivare ad alcuni tra i CEO più potenti del pianeta, sollecitati dal suo sguardo serioso a sostenere una missione epica. Con il tempo, Altman avrebbe imparato che gli individui al potere potevano spianare la strada alle sue ambizioni, proprio come il preside Abbott alle superiori.

Il suo grande progetto scolastico era costruire una versione reale della rete che aveva trovato su AOL. Superati gli impedimenti burocratici dell'istituto, ottenne il benestare del preside alla creazione del primo gruppo di supporto LGBTQ della scuola, una sorta di rete sotterranea tramite la quale gli studenti potevano ricevere consulenza o incontrare loro omologhi. In meno di un anno, aderì oltre una decina di studenti.

Ma per Altman non era abbastanza. Cominciò ad avvicinare gli insegnanti per chiedere loro di applicare adesivi sulle porte delle aule che le segnalassero come spazi sicuri per gli studenti gay. Dopo di che, creò un gruppo (Gay-Straight Alliance) con l'obiettivo di sensibilizzare l'opinione pubblica sui diritti delle persone omosessuali.

Poi decise di fare scalpore in assemblea. Il suo nuovo gruppo entrò di buon'ora in aula magna e posizionò una serie di numeri stampati su tutte le sedie prima che gli altri partecipanti si sedessero. Quando Altman prese la parola al microfono, chiese a tutti coloro che avevano un determinato numero di alzarsi. «Guardatevi intorno», disse al pubblico, mentre una sessantina di ragazzi si alzavano. «Sono uno su dieci di voi. Ecco quante persone in tutta la scuola si identificano come gay.»

Era una dimostrazione audace, ma qualcosa non andò per il verso giusto. Mancavano molti studenti, infatti, tutti iscritti al gruppo cristiano della scuola. Più tardi, Altman venne a sapere che avevano boicottato la presentazione rimanendo a casa o in classe. Indignato dal sabotaggio, tornò ancora una volta nell'ufficio del preside Abbott e chiese che i ragazzi cristiani venissero considerati assenti.

«Non c'era nulla di dannoso nel renderli più consapevoli», argomentò. Non era il tipo da alzare la voce, ma il tono delle parole e l'espressione severa lasciavano trapelare chiaramente la sua rabbia.

«Ho cercato delle giustificazioni», ricorda Abbott, «ma penso che probabilmente avesse ragione.»

Altman lasciò la scuola portando con sé quest'amara lezione: se avevi idee ambiziose, avresti sempre trovato dei detrattori. La soluzione era allearti con chi

aveva potere e autorità e inserirti all'interno di una rete di supporto.

Di lì a poco, Altman fu ammesso alla prestigiosa Stanford University, nel cuore della Silicon Valley californiana, culla di brillanti ingegneri software e imprenditori hi-tech che popolavano di start-up quella regione baciata dal sole. Nonostante l'interesse per la programmazione e il fatto che frequentasse un corso di laurea in informatica, il nostro allampanato diciottenne non riusciva a focalizzarsi su un unico argomento. Era affascinato da tutto, al punto che seguì anche una serie di corsi di discipline umanistiche e scrittura creativa.

Fuori dall'orario dei corsi, poi, guidava per venti minuti verso sud per prendere parte a lezioni che si sarebbero rivelate fondamentali per la sua futura carriera di imprenditore. Passava ore a giocare a poker in un casinò popolare a San Jose, affinando le sue doti di manipolazione e influenza psicologica. Il poker richiede anzitutto la capacità di osservare gli avversari e, talvolta, di depistarli rispetto alla forza della propria mano, e Altman divenne così bravo a bluffare e a leggere i segnali più sottili da riuscire a finanziare la maggior parte delle sue spese universitarie con le vincite. «Lo avrei fatto anche gratis», avrebbe rivelato più avanti in un podcast. «Mi piaceva tantissimo. Lo consiglio vivamente a chiunque voglia imparare qualcosa sul mondo, sugli affari e sulla psicologia.»

Il campo su cui Altman si sarebbe concentrato per trasformare il mondo emerse come parte del programma di studi. Diventò ricercatore presso il laboratorio di AI di Stanford, una nicchia del vasto campus universitario piena di cavi e strani bracci robotici. Il laboratorio era appena stato riaperto e il suo capo era Sebastian Thrun, un informatico dalle idee radicali, con un suadente accento tedesco e penetranti occhi azzurri. Thrun apparteneva a una nuova generazione di accademici che non si accontentava di trascorrere le giornate a scrivere proposte di finanziamento in attesa di una promozione, preferendo collaborare con i giganti della tecnologia. Stanford distava solo otto chilometri dalla sede di Google, e Thrun gestiva anche gli avveniristici progetti «moonshot» di Google X che avevano dato vita alle auto a guida autonoma e agli occhiali per la realtà aumentata.

In classe, Thrun insegnava l'apprendimento automatico, una tecnica che i computer utilizzano per inferire concetti a partire dall'analisi di un'enorme quantità di dati, piuttosto che da una specifica programmazione. Il concetto era fondamentale nel campo dell'AI, anche se il termine «apprendimento» era fuorviante, giacché le macchine non pensano e non apprendono come gli esseri umani. Thrun notò ben presto che quel ragazzo di St. Louis dall'aria seriosa era particolarmente interessato alla possibilità di conseguenze impreviste in ambito AI. Cosa poteva succedere se una macchina imparava a fare la cosa sbagliata?

Thrun spiegò che i sistemi di AI potevano agire nei modi più imprevedibili per raggiungere la loro «funzione di fitness», ovvero l'obiettivo, e se un'AI fosse stata progettata con una funzione di fitness mirata alla sopravvivenza e alla riproduzione, avrebbe potuto involontariamente spazzare via tutta la vita biologica sulla Terra.

Tuttavia, questo non presupponeva che l'AI fosse malvagia. Semplicemente, secondo Thrun, era inconsapevole della portata delle sue azioni. Le sue motivazioni non sarebbero state così diverse da quelle che ci spingono a lavarci le mani. Non è che odiamo i batteri della pelle, e tantomeno vogliamo distruggerli: vogliamo soltanto avere le mani pulite.

Altman rifletté a lungo sull'idea. Da appassionato di fantascienza, si chiese se fosse quello il motivo per cui gli esseri umani non avevano mai avuto contatti con forme di vita aliene. Forse anche gli abitanti di altri pianeti avevano provato a creare un'AI, per poi vedersi annientati dalla loro stessa creatura. Se questo destino si poteva in qualche modo evitare, allora qualcuno doveva progettare un'AI affidabile prima che qualcun altro realizzasse quella pericolosa.

Il seme di quest'idea sarebbe rimasto dormiente nella mente di Altman per poco più di un decennio, prima di sbocciare nella creazione di OpenAI. Al momento, però, era un progetto troppo vasto. Accademici come Thrun costruivano sistemi di AI. Studenti di Stanford come Altman creavano start-up destinate a diventare aziende come Google, Cisco e Yahoo. Il giovane nerd coltivava la stessa ambizione. Gli serviva solo un'idea imprenditoriale. L'illuminazione arrivò al termine di una lezione. «Non sarebbe fantastico se potessi aprire il telefono e vedere una mappa con la posizione di tutti i miei amici?» chiese a Nick Sivo, suo collega di studi nonché amico.

E se avesse creato una mappa digitale per il cellulare su cui fosse stato possibile rintracciare gli amici e ne avesse fatto il prodotto principale di una nuova azienda? Certo, avviare un'azienda non era cosa facile. Bisognava raccogliere fondi da investitori di venture capital e, benché ce ne fossero decine nel raggio di pochi chilometri da Stanford, Altman era comunque giovane e inesperto. La risposta arrivò dall'altra costa degli Stati Uniti, ovvero da Cambridge, nel Massachusetts, dove un imprenditore tecnologico affermato stava avviando quello che aveva definito un *boot camp* per giovani imprenditori. Altman e Sivo decisero di iscriversi al programma di tre mesi, chiamato Y Combinator, e di creare una start-up. Y Combinator sarebbe poi diventato l'acceleratore di start-up di maggior successo di tutti i tempi, finanziando aziende tecnologiche come Airbnb, Stripe e Dropbox, in grado di raggiungere un valore complessivo di 400 miliardi di dollari.

Altman, all'epoca diciannovenne, non sapeva nulla di tutto questo. La maggior parte degli investitori nella Silicon Valley liquidava Y Combinator come una sorta di campo estivo per hacker. Il suo ideatore era Paul Graham, un informatico di quarantun anni che indossava bermuda cargo ed era diventato milionario dopo aver venduto la sua azienda di e-commerce a Yahoo. Raggiunto il benessere economico, Graham si era reinventato come pensatore, pubblicando sul suo sito web saggi su argomenti che andavano ben oltre il campo ristretto degli esperti del software per spaziare dall'economia alla genitorialità, dalla libertà di espressione a cosa significasse essere nerd alle superiori.

Tuttavia, i suoi saggi più popolari, quelli che spingevano ragazzi come Altman a seguirlo con la stessa devozione riservata ai leader spirituali, erano incentrati sulla costruzione di start-up. In quei saggi, Graham sottolineava ripetutamente come la qualità del fondatore di una start-up contasse più di ogni altra cosa. Non serviva un'idea brillante per avviare un'azienda tecnologica di successo. Bastava che ci fosse una persona brillante al comando.

«Il piano di Google, per esempio, consisteva semplicemente nel creare un motore di ricerca che non facesse schifo», scriveva Graham. Ed ecco dove aveva portato questo approccio. I momenti «eureka» appartenevano al passato. Erano i fondatori a fare la differenza, e i migliori erano gli hacker: programmatori disposti a infrangere la saggezza convenzionale per costruire qualcosa di nuovo. In quanto hacker, faceva notare Graham, «potreste essere trentasei volte più produttivi di quanto ci si aspetterebbe da voi in un normale lavoro aziendale».

Avviare un'azienda tecnologica nella Silicon Valley era persino un atto patriottico, poiché ricalcava l'individualismo dei padri fondatori americani. «Gli hacker sono indisciplinati», scriveva. «Questa è l'essenza dell'hacking. Ed è anche l'essenza dell'americanità. Non è un caso che la Silicon Valley si trovi in America e non in Francia, Germania, Inghilterra o Giappone. In quei paesi, la gente colora dentro i margini.»

Il percorso era quasi semplice, per come lo predicava Graham: avvia la tua azienda da zero, inizia con un prodotto minimo funzionante e ottimizzalo nel tempo. Lavora in un ambiente ristretto, perché è meglio avere dieci persone che adorano ciò che hai creato anziché mille che si limitano ad apprezzarlo. E non aver paura di riscrivere le regole lungo il cammino. Anzi, perché non riscrivere addirittura la società nel suo complesso?

Le idee di Graham avrebbero in seguito influenzato profondamente la Silicon Valley, alimentando una nuova convinzione dominante secondo cui la visione dei fondatori di una start-up era così sacra da giustificare una sorta di impunità, quasi fossero esseri divini. È per questo che i fondatori di Google e Facebook hanno potuto crearsi un'immagine di moderni autocrati del business, spesso assicurandosi il controllo della maggioranza delle azioni con diritto di voto e orientando talvolta le rispettive aziende verso scelte discutibili (un esempio emblematico: la scarsa opposizione da parte del consiglio di amministrazione e degli azionisti di Facebook alla strampalata e costosa decisione di Mark Zuckerberg di trasformare l'azienda in un'impresa orientata alla realtà virtuale). Grazie a una struttura azionaria a doppia classe, molti fondatori di start-up tecnologiche, tra cui quelli di Airbnb e Snapchat, hanno potuto mantenere un livello di controllo inusuale sulle loro aziende. Per Graham e altri, questo genere di autorità era più che giustificato. Quando le persone più intelligenti e talentuose avevano una visione a lungo termine, bisognava dar loro la libertà di realizzarla.

In Altman, Graham vedeva quello stesso istinto da hacker: estremamente curioso

e intelligente, era capace di pensare in grande. Ma c'era dell'altro. Quel ragazzino dai capelli scuri e arruffati sembrava a suo agio con le persone più mature, tanto da non avere alcun problema a gestire figure come lo stesso Graham, di vent'anni più grande. Quando Graham gli suggerì di aspettare un anno prima di entrare nel programma di Y Combinator, dato che era appena diciannovenne, il ragazzo rispose che sarebbe venuto comunque. E la simpatia fu immediata.

Gli altri partecipanti a Y Combinator erano per lo più ingegneri e hacker, tra cui i fondatori del popolare forum online Reddit. Nell'ambito del nuovo programma, Graham e sua moglie, Jessica Livingston, assegnarono 6.000 dollari a ogni start-up, una somma basata sull'assegno che il MIT riconosceva durante l'estate a specializzandi e dottorandi. Mentre la maggior parte dei venture capitalist investiva milioni di dollari nelle start-up, Graham incoraggiava i fondatori a fare di più con meno, ovvero a puntare alla cosiddetta «redditività da ramen», scoraggiandoli dal reclutare avvocati, banchieri ed esperti di pubbliche relazioni per portare avanti il lavoro da soli, a costi inferiori.

Lo stesso Graham faceva tutto in economia. Ogni martedì sera preparava la cena – fricassea di pollo, la sua specialità – e invitava gli amici perché parlassero di start-up, mentre la moglie si occupava delle pratiche legali per ogni nuova azienda.

Altman e il suo amico Sivo battezzarono la loro nuova azienda Loopt, e il primo si trasferì a Cambridge, in Massachusetts, per lavorare in quello che allora era l'ufficio di Y Combinator, vicino all'abitazione di Graham. Tra i due si venne a creare un rapporto simile a quello che il ragazzo aveva stretto con il preside delle superiori. Tra i giovani imprenditori di belle speranze, Graham era una specie di leader spirituale, e tutti lo chiamavano semplicemente PG. Altman prendeva molto sul serio gli insegnamenti di Graham e vedeva Loopt non come un'azienda destinata a renderlo ricco, ma come un'idea capace di migliorare il mondo. Lavorava incessantemente al prototipo, tenendosi in piedi a forza di ramen in scatola e gelato al caffè di Starbucks. Lavorò così duramente e mangiò così male da contrarre lo scorbuto a causa della carenza di vitamina C.

Benché fosse un programmatore più che discreto, il giovane Altman era ancora più in gamba come imprenditore. Non aveva remore a contattare i dirigenti di Sprint, Verizon e Boost Mobile per proporre loro una visione ambiziosa: cambiare il modo in cui le persone socializzavano e utilizzavano il proprio telefono. Parlando con voce sommessa e con un'eleganza affinata grazie ai corsi di scrittura creativa, spiegava come Loopt sarebbe diventato essenziale per chiunque possedesse un cellulare. Dato che gli app store ancora non esistevano, ovviamente, bisognava affidarsi agli operatori mobili perché preinstallassero Loopt su alcuni dei primi smartphone. Per questo motivo era fondamentale ottenere il supporto dei dirigenti delle telecomunicazioni, e Altman era un maestro nel promuovere la neonata azienda. Sprint, Verizon, Boost e persino BlackBerry accettarono di integrare il suo servizio sui loro dispositivi.

Alla fine del programma trimestrale di Y Combinator, Altman riuscì a raccogliere fondi con cui far crescere la sua start-up. Presentò la sua idea di Loopt a un gruppo di quindici investitori, per lo più amici facoltosi di Graham, per circa quindici minuti, per poi rivolgersi a investitori con risorse ancora maggiori: le società di venture capital della Silicon Valley. Ricevette diverse offerte e accettò 5 milioni di dollari da due delle principali società di venture capital che in passato avevano finanziato Google, Yahoo e PayPal.

Con tutti questi finanziamenti, Altman lasciò la Stanford University per concentrarsi a tempo pieno su Loopt. Si trasferì a Palo Alto, in California, con un gruppetto di ingegneri che aveva assunto, sistemandosi nello spazio di coworking di Sequoia. Per un certo periodo, il team lavorò ogni giorno fino a tarda sera accanto ai creatori di YouTube, prima che Altman decidesse di traslocare in quello che sarebbe stato il loro primo ufficio a Mountain View, un'area immobiliare di prestigio a pochi isolati dal quartier generale di Google: il cuore pulsante della Silicon Valley.

Questa era la terra dei pensatori visionari. Lì non ci si limitava ad avviare un'azienda: si costruivano imperi. Oppure si puntava a creare qualcosa all'avanguardia, tra scienza e tecnologia. Se volevi condurre ricerche scientifiche su una malattia come l'Alzheimer, potevi rivolgerti a un'università della East Coast o in Europa. Se invece volevi invertire il processo d'invecchiamento, l'unico posto in cui andare era la Silicon Valley.

Il vero punto di forza della regione era la sua rete di contatti. In qualsiasi momento, a un qualsiasi evento, potevi imbatterti in qualcuno in grado di dare una spinta decisiva al tuo business. Fare colazione da Buck's, a Woodside, in California, ti dava la possibilità di incrociare il cofondatore di Yahoo intento a mangiare frutta e yogurt, seduto allo stesso tavolo in cui Elon Musk aveva tenuto il suo primo incontro per ottenere un finanziamento per PayPal. E mentre ordinavi un drink al Musto Bar del Battery Club di San Francisco potevi incontrare uno dei cofondatori di Facebook.

Altman si integrò rapidamente nella rete di programmatori, investitori e dirigenti della Silicon Valley. Se sapevi come inserirti in quel circolo esclusivo, avevi maggiori possibilità di lasciarti travolgere dal successo che trasformava i suoi membri in miliardari. Altman era talmente bravo nel fare networking che riuscì a stringere le giuste connessioni per presentare Loopt alla prestigiosa conferenza annuale per sviluppatori di Apple, nel 2008. Con addosso un paio di jeans e due polo, una verde e una rosa, che gli conferivano l'aspetto di un presentatore di programmi per bambini, il giovane imprenditore dinoccolato dichiarò al suo pubblico che Loopt era il più grande servizio al mondo di social mapping. «Rendiamo possibile la serendipità», disse, fissando la platea con un sorriso appena accennato.

In superficie, sembrava tutto perfetto, ma in realtà Loopt arrancava. La gente non era così entusiasta di usare la mappa digitale di Altman per reperire i propri amici. L'imprenditore dagli occhi sognanti credeva che i ragazzi con uno smartphone desiderassero incontrare fisicamente gli amici tanto quanto lui. Ma organizzare un

rendez-vous con gli amici in un bar o contattare perfetti sconosciuti per una estemporanea partita di basket richiedeva uno sforzo che pochi erano disposti a fare, soprattutto considerando che l'interazione tramite uno schermo era decisamente più semplice. Con il passare degli anni, sempre più persone preferivano interagire con gli amici tramite social network come Facebook che, di fatto, cresceva molto più velocemente di Loopt. Se il primo contava ormai centinaia di milioni di utenti attivi, l'altro si era fermato a cinque milioni di registrazioni.

Inoltre, Loopt stava diventando piuttosto controverso. Un anno dopo aver fondato l'azienda, Altman ricevette una chiamata dall'ex preside delle superiori, Andy Abbott. Gli riferì che i genitori stavano costringendo i figli a usare Loopt per monitorarne gli spostamenti, e un genitore aveva persino chiamato la scuola durante una gita scolastica per segnalare che l'autobus su cui viaggiava il figlio stava andando troppo veloce. «Guarda cos'hai combinato», chiosò, scherzando solo per metà.

Altman aveva sentito di peggio. «Abbiamo ricevuto segnalazioni da gruppi per i diritti delle donne», ammise. Alcuni uomini obbligavano le mogli a installare Loopt per tenerne sempre sott'occhio i movimenti. Un utilizzo inquietante e potenzialmente pericoloso della sua creatura. «Ma stiamo lavorando a una soluzione», si affrettò ad aggiungere. Gli utenti di Loopt avrebbero potuto falsificare la loro posizione. Una donna in difficoltà avrebbe potuto fingersi a casa mentre era al supermercato.

Se molti imprenditori, al suo posto, avrebbero negato il problema, Altman sembrava invece intenzionato ad affrontarlo apertamente. Da adolescente aveva imparato che tenere le cose dentro un cassetto non faceva che peggiorare la situazione. Meglio portarle alla luce. A un certo punto ricevette una chiamata da Jessica Lessin, una tenace reporter sul mondo tecnologico per il «Wall Street Journal», che voleva saperne di più sui problemi di Loopt riguardo alla privacy e su alcune preoccupazioni legate a un possibile uso improprio dell'applicazione. Con sua sorpresa, la giornalista trovò Altman così desideroso di parlare della controversia da inviarle un lungo documento in cui elencava tutti i rischi connessi all'uso della sua app, secondo quanto lei stessa avrebbe riferito in seguito.

Quello che sembrava un suicidio professionale si rivelò una mossa astuta, una sorta di psicologia inversa ben calcolata cui Altman non avrebbe esitato a ricorrere anche in futuro. Mostrandosi più che preoccupato nell'ipotesi degli scenari peggiori, Altman disarmava in partenza ogni critica: non c'era più niente da rinfacciargli, se aveva già fatto tutto da solo. Dava l'impressione di essere fin troppo onesto, anche a suo discapito, benché con tutta probabilità la cosa più giusta da fare con un'app utilizzata per molestare persone vulnerabili sarebbe stata chiuderla.

Alla fine, furono i consumatori a decidere per lui. Altman aveva sottovalutato la scomodità di inviare le proprie coordinate GPS per incontrare qualcuno. «Ho imparato che non puoi costringere gli esseri umani a fare qualcosa che non vogliono fare», avrebbe dichiarato in seguito.

Il giovane imprenditore aveva passato gran parte dei suoi vent'anni ad affannarsi

inutilmente per far crescere Loopt. Aveva sfruttato le nuove notifiche push dell'iPhone per inoltrare avvisi sulle schermate degli utenti e incentivarli a usare la funzione chat di Loopt. Aiutava gli inserzionisti a inviare «offerte lampo» agli utenti di Loopt. Presentava ogni aggiornamento come fosse un successo clamoroso. «La risposta è stata straordinaria», dichiarò in un'intervista del 2010. Ma si trattava di aria fritta. Nel 2012, solo poche migliaia di persone in tutto il mondo usavano regolarmente Loopt. Addio impero. Come la stragrande maggioranza delle start-up tecnologiche, anche Loopt era stata un buco nell'acqua.

Nel mondo delle start-up tecnologiche, l'obiettivo finale di un fondatore è far sì che la sua idea si trasformi in un'azienda da miliardi di dollari o che gli consenta di fare miliardi di dollari vendendola a qualche colosso. Rimanere indipendenti era sempre più difficile, e la maggior parte delle aziende finiva per essere fagocitata da giganti tecnologici come Google o Facebook. Spesso, chi riusciva a vendere utilizzava il ricavato per avviare una nuova azienda e rilanciare il ciclo, da vero e proprio imprenditore seriale. Ma l'uscita di scena di Loopt non fu certo spettacolare. Nel 2012, Altman la vendette a un'azienda di servizi finanziari specializzata in carte prepagate per circa 43 milioni di dollari, cifra appena sufficiente a coprire i debiti verso investitori e dipendenti.

Avrebbe potuto abbandonare la Silicon Valley in quel momento, ma il crollo di Loopt non fece altro che rafforzare la sua determinazione a creare qualcosa di più significativo. Non era il primo cane sciolto nel mondo della tecnologia a trovare ambizioni più grandi nelle ceneri di un fallimento. Circa un decennio prima, Elon Musk era stato cacciato dal consiglio di amministrazione di PayPal. Scottato da quell'esperienza, Musk aveva deciso che lavorare su un progetto superficiale come un servizio di pagamenti per consumatori non facesse più per lui. «La mia prossima azienda dovrà portare un effetto benefico a lungo termine», aveva dichiarato in un'intervista. Pochi anni dopo, avrebbe incontrato i fondatori di Tesla e si sarebbe impegnato a salvare l'umanità dalla minaccia del cambiamento climatico.

Lanciando uno smartphone su un gruppo di persone sedute all'interno dell'esclusivo Battery Club di San Francisco, centrereste almeno tre persone impegnate a salvare il mondo. Molti imprenditori della Silicon Valley si sono convinti, prima o poi, del fatto che la loro app stesse elevando l'umanità; e se alcuni di loro hanno davvero ideato prodotti utili per milioni di persone, molti altri hanno sviluppato un vero e proprio complesso da messia. L'enfasi sull'innovazione ha reso pervasiva questa cultura del salvatore, alimentandola con i principi di Graham secondo cui il fondatore sa sempre meglio degli altri qual è la cosa giusta. Chiunque facesse parte di quella ristretta élite di hacker versati all'innovazione poteva risolvere non soltanto problemi ingegneristici, ma anche dilemmi sociali che da anni affliggevano l'umanità.

Altman voleva che Loopt unisse le persone perché era quello di cui avevano bisogno. Eravamo tutti sempre più incollati agli schermi, impegnati a fare scrolling

senza pensare e a dispensare «like» su diversi social per creare una connessione umana sempre più quantificata. Doveva provare qualcosa di più significativo, dando magari alle persone ciò che non sapevano di volere. Questo era ciò che Apple faceva con successo da anni; nella Silicon Valley, era il segreto che tutti cercavano di decifrare.

Il giovane originario di St. Louis sentiva il bisogno di immergersi nuovamente nel mondo delle start-up e integrarsi così bene nelle reti della Silicon Valley da identificarsi con tutte quelle aziende che proclamavano di voler cambiare il mondo. Si sarebbe trasformato in una versione ancora più profonda dei suoi vecchi mentori, per poi tuffarsi di nuovo su ciò che aveva iniziato a elaborare nel laboratorio di intelligenza artificiale di Stanford. Questo lo avrebbe portato a perseguire un obiettivo ancora più ambizioso: salvare l'umanità da una minaccia esistenziale imminente e portarle in dono una prosperità mai sperimentata prima.

2. VINCERE, VINCERE, VINCERE

Una serie di urla, il rombo di una montagna russa e il clangore di un organetto caratterizzavano l'intro di *Theme Park*, un videogioco del 1994. Un ampio quadrato vuoto di erba pixellata aspettava di essere riempito da bancarelle a forma di enormi hamburger e binari che si alzavano fino ad altezze vertiginose. L'obiettivo era trarre il maggior profitto possibile.

Theme Park non era stato creato da designer di mezza età desiderosi di insegnare i principi aziendali ai bambini, bensì da un adolescente dai capelli scuri di North London di nome Demis Hassabis. Ossessionato dai giochi, Hassabis aveva l'etica del lavoro di un imprenditore della Silicon Valley. Anni prima che diventasse il protagonista di una corsa alla realizzazione dei sistemi di AI più avanzati al mondo, Hassabis stava imparando a gestire un'impresa attraverso la simulazione, secondo quello che sarebbe diventato un tema ricorrente nel suo lavoro e nella sua missione: creare macchine più intelligenti degli esseri umani.

In *Theme Park* si partiva con un saldo di cassa di circa 200.000 dollari per finanziare la costruzione di giostre e pagare gli stipendi del personale. La cifra veniva poi recuperata tramite la vendita di biglietti, di gadget e gelati, e con giochi tipo tiro a segno. Se non assumevi meccanici a sufficienza, le giostre finivano per rompersi; se non c'erano abbastanza guardie di sicurezza, i teppisti invadevano il parco. Se risparmiavi sullo zucchero, i visitatori smettevano di comprare il gelato. Il personale entrava in sciopero e i salari dovevano essere rinegoziati. Benché avesse solo diciassette anni, Hassabis progettò questo delicato equilibrio tra costi e ricompense come una simulazione estremamente sofisticata della gestione aziendale, rendendo il gioco talmente coinvolgente da vendere quindici milioni di copie.

I videogiochi avevano travolto la Gran Bretagna e gli Stati Uniti come un'onda di marea, trascinando i ragazzini in un universo vivido e carico di dopamina, dove le Tartarughe Ninja combattevano in livelli a scorrimento laterale o dove si poteva guidare un pick-up su spericolate piste sterrate. Ma Hassabis pensava che i migliori videogiochi fossero simulazioni che funzionavano come microcosmi della vita reale. I cosiddetti «God games» ti permettevano di esercitare il potere di creazione e distruzione. Invece di controllare un singolo personaggio come Mario, modellavi le vite di migliaia di personaggi virtuali, creando paesaggi o dirigendo i progressi di un'intera civiltà. Potevi erigere una città e poi colpirla con un disastro naturale, oppure riempire un parco tematico con centinaia di visitatori.

Questa tecnologia poteva essere usata per divertirsi, ma anche per imparare tutta una serie di cose, da come gestire un'impresa ai misteri dell'universo. Al di là del valore legato all'intrattenimento, Hassabis sarebbe stato presto travolto dal desiderio di utilizzare i giochi per creare un'intelligenza artificiale così sopraffina da aiutarlo a svelare i segreti della coscienza umana.

Questa vocazione a comprendere i misteri dell'universo andava oltre gli obiettivi della maggior parte degli altri scienziati, ed è un aspetto che può sembrare strano se non si tiene conto che lo stesso Hassabis era cresciuto come un enigma, un genio matematico solitario in una famiglia di creativi bohémien. Sua madre, Angela, era una devota battista emigrata nel Regno Unito da Singapore. A North London, presso la famiglia ospitante da cui alloggiava, aveva conosciuto il futuro marito, un greco-cipriota dallo spirito libero di nome Costas Hassabis. Per quanto apparentemente incompatibili come calzini e sandali, i due si erano sposati e avevano avuto tre figli, il primo dei quali era appunto Demis. Costas saltava da un lavoro all'altro, come insegnante e gestore di un negozio di giocattoli, e prima che Demis compisse dodici anni aveva già costretto la famiglia a dieci traslochi.

A quel punto, era chiaro che Demis era diverso dagli altri bambini. A quattro anni aveva già battuto il padre e lo zio a scacchi, e a sei sconfiggeva la maggior parte dei coetanei nei campionati locali, appollaiato su un cuscino o su un elenco telefonico per riuscire a vedere meglio la scacchiera. Divorava libri ed era curioso di tutto, ma canalizzava gran parte delle sue energie mentali nei giochi. Quando suo padre portava a casa giochi da tavolo con pezzi mancanti, li usava per inventarne di nuovi e giocarci con il fratello e la sorella minori.

Ma il vero divertimento doveva ancora arrivare. Un decennio prima che Sam Altman scoprisse le chat room di AOL, Hassabis si immergeva a capofitto in una tecnologia ancora più rudimentale, fatta di grezzi pixel su schermi neri. Nel 1984, a otto anni, con i soldi vinti in un torneo di scacchi acquistò uno ZX Spectrum 48. Si trattava di uno dei primi personal computer e consisteva in una tastiera nera e spessa che si collegava alla TV e impiegava cassette a nastro per riprodurre grafiche colorate sullo schermo.

Dopo aver comprato alcuni libri di programmazione, Hassabis imparò da solo a realizzare giochi per lo Spectrum. Prima di andare a letto, impostava dei calcoli da eseguire durante la notte. La mattina dopo, questi erano completati. Per Hassabis, fu una rivelazione: aveva delegato il suo lavoro cognitivo allo Spectrum. Il computer operava come un'estensione della sua mente.

Appassionatosi alla programmazione, allora un mondo di nicchia, passò a un più potente Commodore Amiga 500, un ingombrante set di dispositivi bianchi che includeva un mouse e il monitor, e con alcuni compagni di scuola fondò un club di hacking. Insieme, scrivevano linee di codice per creare frammenti colorati sullo schermo, riproducendo scene di videogame a cui avevano giocato. In tutto questo, Hassabis puntava sempre a rendere i progetti via via più complessi. Smontava il

computer e poi lo rimontava. Creò un gioco di scacchi digitale e convinse il fratello minore, George, a provarlo.

Gli scacchi rimanevano il fulcro della sua vita, dal momento che sognava di diventare campione del mondo. Sostenendo tali aspirazioni, sua madre iniziò a fargli homeschooling, così da permettergli di dedicare più tempo allo studio del gioco. Durante le vacanze scolastiche, viaggiava per partecipare a tornei di scacchi, allenandosi a ritmi incessanti. Anni dopo, avrebbe detto che i giochi erano come una palestra per la mente e che gli scacchi erano l'allenamento definitivo. Così come il poker avrebbe insegnato a Sam Altman la psicologia e gli affari, gli scacchi insegnarono a Hassabis a elaborare strategie partendo dalla fine: visualizzavi un obiettivo e procedevi a ritroso per raggiungerlo.

Ma tutto cambiò quando Hassabis, a undici anni, partecipò a un campionato di scacchi in Liechtenstein. Stava giocando contro il campione nazionale danese, e la partita era diventata una vera e propria maratona. Dopo dieci ore di gioco, con le menti ormai appannate, il danese cercò di forzare un pareggio. Il giovane Hassabis aveva ancora re e regina, ma l'avversario era in vantaggio di una torre, un alfiere e un cavallo. Ormai sfibrato e temendo di subire uno scacco matto, Hassabis si arrese.

Il campione danese rimase sbalordito. «Perché ti sei arreso?» gli chiese.

Poi gli mostrò la mossa difensiva che avrebbe potuto fare. Hassabis fissò la scacchiera. A volte, il fallimento scatena ambizioni ancora più grandi. Se non riesci a tollerare la sconfitta, inseguire un obiettivo più ambizioso può costituire una forma di consolazione, e Hassabis aveva appena fallito dopo uno sforzo titanico. Guardandosi intorno, mentre gli altri cervelloni se ne stavano curvi sulle loro scacchiere, i neuroni in piena attività, all'improvviso gli venne da pensare che quel torneo non era che uno spreco di intelligenza. Lì dentro c'erano alcuni dei migliori pensatori strategici del pianeta. E se si fossero messi a lavorare per risolvere problemi più vasti? Era ormai il secondo miglior giocatore di scacchi under quattordici al mondo, ma tutto questo rimaneva pur sempre un gioco.

Hassabis disse ai suoi genitori che non voleva più partecipare ai tornei di scacchi e tornò tra i banchi di scuola. Era un ragazzo tranquillo e sensibile, che ascoltava Enya e aveva imparato da solo a suonare al piano *Watermark*. Il suo film preferito era *Blade Runner*, un classico della fantascienza con protagonista un detective che dà la caccia a replicanti AI ribelli quasi indistinguibili dagli esseri umani. Coinvolto emotivamente dai passaggi più toccanti del film, ascoltava a ripetizione la struggente colonna sonora di Vangelis durante la scena finale, quando il cattivo si lamenta del fatto che i suoi ricordi andranno «perduti nel tempo, come lacrime nella pioggia».

La domenica, sua madre portava regolarmente Demis e i suoi fratelli alla Hendon Baptist Church, nella zona nord di Londra, un'imponente struttura in pietra grigia che dominava i sobborghi dall'alto di una collina. Era una piccola enclave internazionale, con fedeli provenienti da paesi come le Filippine, il Ghana, la Francia e l'India. Per un ragazzo metà cipriota e metà singaporeano come Hassabis, era un ambiente in cui

per una volta poteva sentirsi integrato. Le domeniche, lì, offrivano un'alternativa più vivace alle funzioni ingessate della chiesa anglicana. La gente alzava le mani al cielo e cantava al ritmo della batteria e della band. Il pastore tralasciava i dogmi per sottolineare l'importanza del rispetto reciproco. Le preghiere erano cariche di emozione e la chiesa non aveva remore a mostrarsi apertamente evangelica.

I battisti erano una piccola minoranza, nel Regno Unito, pur essendo una delle denominazioni cristiane più numerose negli Stati Uniti. Contavano circa 150.000 membri attivi, a fronte del milione di praticanti della chiesa anglicana. Ma la religione e il concetto stesso di Dio affascinavano Hassabis, tanto da spingerlo a chiedersi se fosse possibile individuare questo Dio attraverso la scienza. A sedici anni, terminate le superiori con due anni di anticipo, lesse *Il sogno dell'unità dell'universo* del Nobel per la fisica Steven Weinberg. Il libro trattava della grande, quasi donchisciottesca ricerca di una teoria unificata della natura. Weinberg credeva che ci potesse essere un modo per spiegare tutte le forze fondamentali dell'universo attraverso un unico insieme di equazioni, un po' come l'equazione di Einstein $E = mc^2$ riassumeva la relazione tra energia e massa. Idealmente, questa «teoria del tutto» doveva essere così sintetica da rientrare in una singola pagina o addirittura in un'unica equazione.

Per Hassabis, il fatto che gli scienziati non stessero facendo molti progressi nell'individuare questo quadro teorico era intrigante e, allo stesso tempo, lo lasciava sbalordito. Avevano bisogno di aiuto, pensò. Avevano bisogno di una potenza intellettuale maggiore. Avrebbe potuto aiutarli in qualche modo? Gli venne in mente il pesante Commodore Amiga che eseguiva calcoli durante la notte mentre lui dormiva. Forse un computer più intelligente avrebbe fatto al caso loro. Se fosse riuscito a rendere i computer più intelligenti, un'estensione più performante della mente umana, forse avrebbe potuto aiutare gli scienziati a trovare una risposta a quelle domande complicate sull'universo, fino a scoprire, magari, un'origine divina.

«Sembrava la soluzione perfetta, una mega-soluzione», spiegò più avanti Hassabis in un'intervista con il giornalista del «New York Times» Ezra Klein. Aveva pensato di studiare fisica, all'università, ma dopo aver letto il libro di Weinberg capì di dover puntare a qualcosa di più grande. Se avesse studiato informatica e il nascente campo dell'intelligenza artificiale, avrebbe potuto costruire lo strumento scientifico definitivo e fare scoperte che migliorassero la condizione umana. Incapace di liberarsi della sua fissazione per i giochi, Hassabis delineò un piano a lungo termine per combinare entrambi i suoi interessi. La chiave era concentrarsi su giochi che simulassero il mondo reale. Negli anni Ottanta, i giochi riuscivano già a simulare le caratteristiche rudimentali di un'intera civiltà. Se un computer era in grado di replicare tutto un mondo nei dettagli, forse un dispositivo più intelligente avrebbe potuto anche correggere i difetti del mondo reale: capire cosa fare nella simulazione e applicarlo alla vita reale.

Hassabis trasse ispirazione dai «God games», in particolar modo da *Populous*. «Quello che mi affascinava di questi giochi era che erano mondi viventi, che si

evolvevano a seconda di come giocavi», avrebbe raccontato. «Potevi simulare una qualsiasi parte del mondo e sperimentarci su.»

La grafica retrò e pixellata del gioco nascondeva la sua complessità. Il giocatore assumeva i panni della divinità di una verde vallata punteggiata di case, e con i suoi poteri guidava gli abitanti – suoi seguaci – in battaglia contro i fedeli di altre divinità. Poteva aumentare e abbassare la pendenza del terreno, appiattendolo in modo da consentire ai suoi di costruire case e moltiplicarsi. Poteva anche scatenare terremoti. *Populous* fu il precursore del genere, e Hassabis lo amava a tal punto da maturare la determinazione a lavorare per l'azienda che lo aveva ideato, la Bullfrog Productions. Tuttavia, quando partecipò a una competizione che metteva in palio un posto in azienda, il giovane Hassabis perse. Così, prese il telefono e contattò direttamente la Bullfrog, chiedendo di poter fare un'esperienza lavorativa di una settimana. L'azienda accettò e rimase così colpita da offrirgli un impiego estivo anche se aveva solo quindici anni.

L'anno dopo, quando ottenne un posto a Cambridge per seguire un corso di studi in informatica, l'università gli fece sapere che era troppo giovane e che doveva aspettare almeno un anno. Così, trascorse quel periodo sempre alla Bullfrog, con un compenso in contanti e un alloggio presso l'ostello della gioventù vicino agli uffici dell'azienda a Guildford, nel Surrey. Dopo gli inizi come tester di videogiochi, Hassabis fu promosso a level designer sotto la guida diretta del fondatore dell'azienda, Peter Molyneux.

Con la testa rasata e le sue polo nere, Molyneux sembrava più il proprietario di un pub che una leggenda del mondo videoludico. Nel corso degli anni, era diventato anche una figura controversa nell'industria per via dell'abitudine ben documentata di lasciarsi andare a promesse eccessive su ogni nuovo progetto. Faceva affermazioni ambiziose sulle meccaniche o sulle caratteristiche dei giochi, dicendo, per esempio, che i giocatori avrebbero potuto piantare una ghianda nel mondo virtuale di *Fable* e qualche giorno dopo trovarvi un albero. Cosa assolutamente non vera.

Ma Molyneux aveva anche grandi idee capaci di promuovere l'industria e, al momento, godeva del successo di *Populous*. Ai suoi occhi, Hassabis era un ragazzino insolitamente curioso, quasi precoce. Il giovane prodigio lo tempestava di domande sui limiti tecnici dei giochi della Bullfrog e chiedeva perché venissero definite come «intelligenza artificiale» certe caratteristiche che sembravano sistemi software di base.

«Nessun incarico, per quanto assurdamente gravoso, gli appariva come un ostacolo», avrebbe ricordato Molyneux. Quando gli altri membri del team rimbalzarono la sua proposta di creare un gioco di simulazione su un parco a tema, dichiarandosi più interessati a giochi con spade e combattimenti, Hassabis si offrì volontario per costruire quello che sarebbe diventato il vivace mondo di montagne russe e bancarelle di *Theme Park*. Mentre i due progettavano insieme il nuovo gioco, con il supporto di un paio di artisti digitali, Molyneux divenne una sorta di mentore

per Hassabis. Tra una sessione di programmazione e una riunione sul gameplay, si finiva spesso per discutere delle possibilità dell'intelligenza artificiale. Secondo Hassabis, mancava probabilmente una decina di anni perché l'AI diventasse senziente e sopravanzasse gli esseri umani.

«Il futuro dell'AI sembrava quasi a portata di mano», ricorda a distanza di anni il senior game designer. «L'altra domanda su cui filosofavamo spesso era: "Perché gli esseri umani dovrebbero essere gli unici a creare cose?". Perché non affidare il peso della creatività a un'AI?» Prefiguravano un futuro in cui l'AI avrebbe scritto musica e poesia, e persino progettato giochi.

Per il momento, però, i sistemi che utilizzavano per conferire a *Theme Park* un tocco di realismo rientravano a stento nella definizione di intelligenza artificiale. Tecniche di apprendimento automatico permettevano di assegnare personalità diverse ai personaggi: alcuni visitatori erano più impulsivi e propensi a spendere, mentre altri erano più parsimoniosi. Il gioco fu un successo. Se *Populous* aveva venduto cinque milioni di copie, *Theme Park* riuscì a venderne tre volte tanto.

Questo rese Hassabis una sorta di celebrità, quando infine arrivò a Cambridge. Prese in prestito la Porsche 911 di Molyneux e la guidò su e giù per il campus, desideroso di impressionare gli studenti. Dopo aver trascorso le vacanze precedenti partecipando a tornei di scacchi, affrontò il suo primo anno all'università come se fosse una seconda vacanza, uscendo la sera con i compagni di corso per poi restarsene a letto al mattino, ad ascoltare i Prodigy mentre il sole filtrava dalle finestre. Quando non beveva vino al bar del college fino a diventare paonazzo, giocava partite lampo a scacchi o gareggiava in corse clandestine con la Porsche presa in prestito. Alla fine, si schiantò e dovette chiamare il suo mentore per scusarsi. «Era la seconda volta che la distruggeva», ricorda Molyneux con una smorfia. Ma era difficile arrabbiarsi con quel ragazzo prodigio dal sorriso contagioso. «È davvero irresistibile.»

A Cambridge, Hassabis conobbe i membri del suo futuro circolo intimo, tra cui Ben Coppin – un altro studente di informatica destinato a diventare il responsabile dello sviluppo del prodotto in DeepMind –, con cui parlava di religione e di come l'AI avrebbe potuto risolvere problemi di portata globale. Ma DeepMind distava ancora più di dieci anni. Il piano era laurearsi e tornare a lavorare subito con Molyneux. A quel punto, però, Hassabis si imbatté nella più bizzarra delle candidature di lavoro. Un giorno, infatti, ricevette una bottiglia con dentro una lettera macchiata di tè e con i bordi bruciati. L'autore della missiva, scritta a mano in bella grafia, raccontava di essere naufragato su un'isola chiamata Korporate. Hassabis colse il messaggio al volo, perché anche lui avrebbe detestato l'idea di lavorare per una gigantesca corporation.

Il mittente era Joe McDonagh, un programmatore che lavorava per un vasto conglomerato chiamato British Telecom ed era un grande appassionato di videogiochi. Joe sognava disperatamente di lavorare per un'azienda di videogiochi e,

per la sua immensa gioia, fu invitato a presentarsi a casa di Molyneux per un colloquio. Ad aprirgli fu un giovane dall'aspetto quasi elfico, con la barbetta sul mento e un casco di capelli neri. «Ricordo di aver pensato: "E questo chi è?"» racconterà McDonagh. Hassabis, che lo aveva accolto nelle vesti di dirigente dell'azienda, dimostrava anche meno dei suoi ventun anni.

McDonagh si rese subito conto di trovarsi davanti a un altro appassionato di videogiochi estremamente competitivo. Quando tirò in ballo la sua passione per l'origami, Hassabis lo sfidò a chi realizzava per primo una gru di carta e lo bruciò sul tempo. Dopo di che, trascorsero il resto del pomeriggio davanti a diversi giochi da tavolo. In seguito, quando McDonagh richiamò per chiedere notizie sul lavoro, gli fu detto che lo strano ragazzo con cui aveva sostenuto il colloquio aveva lasciato l'azienda. Hassabis se n'era andato perché il suo mentore non riusciva più a tenere il passo con le sue ambizioni tecniche. «Non eravamo abbastanza veloci, per lui», ammette lo stesso Molyneux.

McDonagh riuscì a rintracciare il numero di telefono di Hassabis e lo chiamò per scoprire cosa fosse accaduto. «Sto mettendo su una nuova azienda», gli spiegò l'altro. Si sarebbe chiamata Elixir Studios e prevedeva un'AI all'avanguardia al centro di un «God game» che mirava a simulare il mondo.

Era un progetto incredibilmente ambizioso, ma McDonagh non si tirò indietro. Entrò in Elixir come responsabile della progettazione, un compito che consisteva nell'immaginare nuovi e straordinari mondi. Avendo imparato l'arte del cosiddetto «marketing bombastico» dal suo ex mentore, Hassabis non si mostrava affatto timido o modesto riguardo ai suoi obiettivi davanti alla stampa. Apparve sulla copertura di «Edge», rivista di riferimento dei videogiochi negli anni Novanta, vantandosi di voler creare giochi di un livello tale da sollevare la categoria dal suo mercato di nicchia per adolescenti. Avrebbe creato qualcosa di così brillante da stimolare la curiosità persino dei lettori dell'«Economist». «Volevo dimostrare che i giochi avevano la stessa dignità di film e libri», racconta Hassabis. Tracciò dunque un piano a lungo termine. Una volta raggiunto il successo, avrebbe venduto Elixir per fondare un'azienda interamente incentrata sull'intelligenza artificiale.

Si concentrò sulla creazione di un gioco chiamato *Republic: The Revolution*, una simulazione politica in cui i giocatori dovevano rovesciare il governo di un paese totalitario fittizio dell'Europa orientale. Hassabis voleva il massimo del realismo. McDonagh trascorse ore alla British Library studiando la storia dell'Unione Sovietica per supportare il suo giovane capo nella creazione di una storia realistica. Hassabis lavorava più sul comparto tecnico, supervisionando la realizzazione di un'intelligenza artificiale in grado di gestire un milione di persone virtuali, un obiettivo decisamente ambizioso, considerando che fino a quel momento il limite per la maggior parte dei «God games» era prossimo ai duemila personaggi virtuali. Voleva che i giocatori potessero zoomare dall'immagine satellitare di una città fino a un petalo sul balcone di un palazzo.

L’ex campione di scacchi assunse i programmatori più brillanti che riuscì a reperire, molti dei quali laureati a Oxford e Cambridge. Rafforzò lo spirito di squadra incoraggiandoli a giocare in qualsiasi momento. Lui era un asso in tutti i giochi, dal videogame *Starcraft* al gioco da tavolo strategico *Diplomacy*. Il calciobalilla era uno sport cruento nel quale Hassabis sfoderava il suo «colpo della vipera» per sparare la palla verso la porta con tutta la forza che aveva in corpo. Sul campo di calcio reale, giocava come attaccante nella squadra di Elixir, composta da ingegneri informatici non proprio in forma. Quando giocavano in cinque contro cinque con un gruppo di ragazzi di North London, alti il doppio di lui, Hassabis aggrediva pallone e tibie come un terrier arrabbiato, segnando un sacco di gol.

Con l’avvicinarsi della deadline per l’uscita del gioco, il fondatore e i suoi programmatori presero a lavorare dalle 10 del mattino alle 6 del mattino dopo, concedendosi appena tre o quattro ore di sonno in sala riunioni, e finendo a volte per russare sulla scrivania, con i gamepad penzolanti dalle mani. Niente più uscite serali, e lo stesso Hassabis si impose di non bere più, per non compromettere in alcun modo le sue capacità cerebrali.

Le sue ambizioni per la grafica e la tecnologia AI di *Republic* rasentavano l’assurdo, dal momento che stava sviluppando un sistema di gran lunga eccedente le capacità computazionali dell’epoca. Ma non aveva senso costruire un paese virtuale se non riusciva a popolarlo di migliaia di personaggi reali e credibili. «Non volevo che le persone fossero semplici punti astratti che vagavano a caso sullo schermo», raccontò alla rivista «Edge». «Volevo mariti, studenti, casalinghe e ubriaconi, ognuno con una vita tutta sua e plausibile.»

Non c’era modo migliore di mostrare le straordinarie capacità dell’AI che attraverso un gioco. A quel tempo, le ricerche più avanzate in materia si stavano sviluppando proprio nell’industria videoludica, dove software sempre più sofisticati contribuivano a plasmare mondi vividi e un nuovo stile chiamato «emergent gameplay». Invece di seguire un percorso prestabilito come in *Super Mario Bros.*, il giocatore veniva catapultato in un mondo virtuale, dotato di alcuni strumenti e lasciato libero di cavarsela da solo. Questa era l’essenza di *Grand Theft Auto* e, in seguito, anche quella di *Minecraft*, il videogioco più venduto della storia.

Benché Hassabis fosse convinto di trovarsi sulla strada giusta, c’era un problema: *Republic: The Revolution* era noioso, e non c’era trappola peggiore per un progettista di videogiochi. Il team aveva dedicato quattro dei cinque anni di sviluppo all’aspetto tecnologico, trascurando però di perfezionare il gameplay. Creare un grande videogioco richiede iterazioni costanti. In genere, si parte da un prodotto grezzo ma giocabile e poi si testa migliaia di volte allo scopo di migliorarlo. Ma i creatori di Elixir non avevano potuto prendersi il tempo necessario per rendere il gioco più interessante, perché le aspettative del loro capo sull’aspetto tecnologico erano troppo elevate.

«L’essenza di un videogioco è il senso di immersione e la sensazione che

trasmette», ricorda McDonagh. «*Republic* non aveva né l'una né l'altra cosa. Ci siamo ritrovati in un buco nero tecnologico.» I programmatori di Elixir sapevano che il gioco non era all'altezza delle aspettative. All'uscita, i critici confermarono i loro peggiori timori, assegnandogli recensioni contrastanti e giudicandolo troppo complesso. Le vendite furono modeste.

«Era troppo ambizioso per l'epoca», ammette Hassabis. «Avevo fretta di lasciare il segno sotto il profilo tecnico e artistico.»

Ciò non gli impedì di riprovarci ancora, lanciando un nuovo gioco altrettanto ambizioso, *Evil Genius*, in cui i giocatori vestivano i panni di un cattivo in stile James Bond con l'obiettivo di conquistare il mondo. Il gioco aveva un umorismo intelligente e autoironico, ma faticò ugualmente a raggiungere un successo di massa. Nel cercare di migliorare la situazione con *Evil Genius 2*, Hassabis si trovò a dover fare i conti con i costi spropositati di tutti gli investimenti tecnologici. Nel 2005, il ragazzo brillante che voleva andare oltre i limiti del gaming fu costretto a chiudere Elixir. Quell'esperienza fu devastante per Hassabis, che dalla scacchiera al calciobalilla, passando per la scuola, era sempre stato abituato a vincere.

Il contraccolpo dell'industria videoludica britannica rese l'umiliazione ancora più cocente. Hassabis aveva saputo far crescere magistralmente l'entusiasmo intorno a Elixir, presentandola alla stampa e al settore come una realtà in grado di rivoluzionare il mondo dei videogiochi grazie a una tecnologia innovativa. McDonagh ricorda di aver partecipato a una conferenza e di aver menzionato il suo lavoro per Elixir. Un pezzo grosso dell'industria videoludica britannica, sentendolo parlare, reagì con una risata di scherno, e lui si sentì morire dentro. «Il fallimento fu incredibilmente doloroso.»

A un certo punto, McDonagh e Hassabis ebbero un duro confronto sull'inevitabile tracollo dell'azienda, la prima e unica volta in cui il primo sentì l'altro, normalmente così pacato, alzare la voce. «È stato un incubo», racconta. «Eravamo tutti di Oxford e Cambridge. Tutti abituati a vincere, vincere, vincere. Non avevamo mai perso, prima, e quella volta lo abbiamo fatto in modo clamoroso.»

Nel suo desiderio di mostrare al mondo le meraviglie dell'AI, Hassabis aveva commesso un errore cruciale: aveva cercato di costruire un gioco intorno alla sua vera ossessione. Se voleva creare macchine più intelligenti degli esseri umani, doveva ribaltare l'approccio. Non doveva più ricorrere all'intelligenza artificiale allo scopo di realizzare un gioco eccezionale, bensì usare i giochi per sviluppare a un'AI eccezionale.

Anni dopo, ormai sulla trentina, McDonagh si ritrovò di nuovo al telefono con il suo vecchio capo per valutare un'altra offerta di lavoro. «Sto fondando quest'azienda chiamata DeepMind», disse l'infaticabile imprenditore, evidentemente galvanizzato dalle sfide impossibili.

"Non posso ricascarci", pensò McDonagh. Quindi rispose: «No».

Poi assistette stupefatto alla nuova impresa di Hassabis che, individuato un altro

obiettivo incredibilmente ambizioso, questa volta riuscì persino a superare ogni aspettativa, sviluppando quello che sembrava il sistema di intelligenza artificiale più avanzato al mondo, almeno fino all’arrivo di Sam Altman.

3. SALVARE GLI ESSERI UMANI

In una calda giornata estiva del 2006, Altman era sdraiato sul pavimento del suo monolocale a Mountain View, in California, con addosso solo un paio di pantaloncini da palestra. Le braccia distese, cercava di respirare. Era a metà di un estenuante weekend dedicato alla negoziazione di un accordo per Loopt, e le prospettive non erano rosee. Nell'appartamento c'erano trentacinque gradi. Altman si sentiva esplodere per lo stress, come avrebbe raccontato nel 2022 al podcast *Art of Accomplishment*.

Per anni, si era ripetuto che per un imprenditore era normale. «È così che ci si sente», pensava. «Ma saperlo non è di alcun aiuto.» Lo stress non faceva che peggiorare le cose.

Il fallimento di Loopt gli aveva insegnato che non poteva costringere le persone a fare ciò che non volevano. Ma gli aveva anche impartito una lezione più intima: come disimpegnarsi emotivamente dalle situazioni complicate. Quel momento sul pavimento divenne un punto di svolta. Avrebbe vissuto diversamente. Con più distacco: era quello, il trucco.

Dopo aver venduto Loopt e rotto con Nick Sivo (che era stato a lungo anche suo compagno di vita), e dopo aver lavorato per un po' per l'azienda che aveva rilevato la sua, Altman si prese un anno per fare ciò che gli piaceva. Si disimpegnò completamente. In genere, nella cultura *hustle* della Silicon Valley, l'idea di prendersi un anno sabbatico non è vista di buon occhio, e Altman se ne accorse immediatamente. Se iniziava a parlarne con qualcuno, a una festa, gli occhi del suo interlocutore correvano subito a cercare qualcun altro con cui conversare.

Mantenne un filo di connessione con la Bay Area della California, lavorando come partner part-time per Y Combinator. A quel punto, gli investitori di venture capital avevano cambiato idea sul programma, che ormai vedevano come una vera e propria fabbrica di aziende internet di alta qualità. Alcune delle start-up nate lì erano diventate realtà solide come Reddit e Scribd. Per i fondatori di start-up, «YC» era ormai una porta di accesso alla Silicon Valley. Migliaia di imprenditori tech facevano domanda ogni anno, ma solo un centinaio venivano accettati.

Per il resto del suo anno sabbatico autoimposto, Altman si dedicò a una vasta gamma di interessi, leggendo decine di libri sugli argomenti più disparati, dall'ingegneria nucleare alla biologia sintetica, dagli investimenti all'intelligenza artificiale. Visitò altri paesi, soggiornò in ostelli, partecipò a conferenze e investì in

diverse start-up con una parte dei 5 milioni di dollari circa guadagnati dalla vendita di Loopt.

Ammetteva pubblicamente che quasi tutte le aziende in cui aveva investito erano fallite, ma era convinto che questo gli servisse ad allenare il «muscolo» capace di identificare i progetti con maggiore probabilità di successo. Sbagliare spesso andava bene, purché ogni tanto si facesse centro «in grande», investendo per esempio in una start-up destinata a fare il botto e garantendosi, così, un'uscita spettacolare.

Se vivere la vita era come dipingere, allora Altman utilizzava il rullo più grande a disposizione per coprire un'area quanto più ampia possibile. Tuttavia, si sentiva sempre più attratto dall'intelligenza artificiale. Più o meno nel periodo in cui vendette Loopt, fece un'escursione con alcuni amici del settore tecnologico e si ritrovò a discutere del futuro della ricerca sull'intelligenza artificiale, come riportato in un articolo sul «New Yorker». Altman era dell'idea che, a un certo punto, durante la sua vita, gli hardware sempre più potenti e i sistemi di apprendimento automatico sempre più avanzati sarebbero stati probabilmente in grado di replicare il funzionamento del cervello umano.

Questa intuizione gli rivelò qualcosa di fondamentale sul ruolo dell'umanità al vertice della catena alimentare. Se la nostra intelligenza poteva essere simulata da un computer, eravamo davvero così unici? La sua risposta a questa domanda era negativa e, benché all'inizio suonasse come una constatazione deprimente, riuscì presto a ribaltarla in un punto su cui fare leva. Se gli esseri umani non erano così speciali, allora potevano essere replicati dai computer, e persino migliorati. E forse poteva essere proprio *lui* a fare in modo che ciò avvenisse.

Per molti versi, Altman si stava facendo interprete della mentalità tipica della Silicon Valley, che vedeva la vita stessa come un enigma ingegneristico. Si potevano risolvere problemi di enorme portata applicando gli stessi passaggi impiegati per ottimizzare un'app. Quest'orientamento derivava in parte dalla formazione tipica degli ingegneri, abituati ad affrontare i problemi tecnici con un approccio sistematico e logico, secondo un metodo radicato a fondo nella loro cultura. Il successo si misurava in base all'efficienza del proprio software. Era naturale che queste tecniche così apprezzate si estendessero ad altri aspetti della società e della vita.

Non sorprende che Altman amasse usare il linguaggio informatico per riferirsi agli esseri umani, come quando in un'intervista a una rivista dichiarò che «apprendiamo solo due bit al secondo». Il bit è l'unità base dell'informazione, generalmente rappresentata come 0 o 1 nel codice binario, e Altman se ne servì come metafora per sottolineare quanto fosse limitata la nostra capacità di elaborare informazioni. Confrontando i meccanismi del cervello umano con il funzionamento di un computer, quest'ultimo risultava enormemente più rapido nell'elaborare i bit, operando a velocità misurabili in gigabit o terabit al secondo.

Se Altman voleva realizzare una macchina in grado di superare l'intelligenza umana, non c'era dubbio sul fatto che dovesse rimanere nella Silicon Valley, il posto

in cui si progettava il domani.

«Qui c'è una convinzione incrollabile nel futuro», disse una volta. «Ci sono persone che prenderanno sul serio le tue idee più stravaganti, invece di deriderti.» La Silicon Valley offriva anche una rete di contatti in continua espansione dove vigeva il principio del *do ut des*: se aiutavi qualcuno a raccogliere fondi per la sua start-up, magari lui ti avrebbe aiutato a reclutare un ingegnere di talento.

Rientrato dai suoi viaggi, Altman avviò una società di finanziamento per start-up emergenti chiamata Hydrazine Capital, per puntare su realtà che si occupavano degli ambiti più disparati, dalle scienze della vita ai software per l'istruzione. A tal fine, fece leva sui suoi contatti con alcuni dei finanziatori più influenti della Silicon Valley. Paul Graham e Peter Thiel, uno dei primi investitori in Facebook, contribuirono al fondo da 21 milioni di dollari raccolto alla fine da Altman. Il secondo, oggi un enigmatico miliardario con idee che sfiorano la fantascienza, sarebbe diventato una figura chiave nella corsa allo sviluppo di intelligenze artificiali potenti, finanziando sia Altman sia Hassabis a Londra. Thiel faceva parte della cosiddetta «PayPal Mafia», un gotha di cofondatori e dirigenti del gigante dei pagamenti online che per anni avevano investito ciascuno nelle aziende degli altri e che includeva anche Elon Musk e il fondatore di LinkedIn Reid Hoffman.

Altman destinò circa il 75 per cento dei fondi di Hydrazine alle aziende che uscivano da Y Combinator, una strategia che si rivelò vincente. Nel giro di quattro anni circa, il valore del fondo aumentò di dieci volte, grazie agli investimenti in start-up che appartenevano alla sua rete di contatti in espansione, molti dei quali parte dell'élite della Silicon Valley.

Il fondo investì in Reddit, la start-up della sua prima classe di YC, e in Asana, l'azienda di software per imprese fondata dal miliardario cofondatore di Facebook, Dustin Moskovitz. Negli anni successivi, entrambe queste relazioni si sarebbero rivelate cruciali nella creazione di un'AI ultrapotente.

Altman era consapevole del fatto che le ricompense finanziarie immediate non erano così preziose nel lungo periodo quanto le connessioni personali. Ecco perché si sentiva a disagio con il comportamento assertivo che doveva adottare nei confronti degli imprenditori in qualità di venture capitalist. Era un lavoro in cui bisognava ottenere la maggior quota di partecipazione investendo meno denaro possibile. Altman trovava anche un po' sgradevole l'ossessione della Silicon Valley per la ricchezza estrema. Era più interessato alla gloria che derivava dal mettere in piedi progetti entusiasmanti. Tra un investimento e l'altro, ridusse i suoi beni materiali a una casa con quattro camere da letto a San Francisco, una proprietà a Big Sur, in California, e 10 milioni di dollari in contanti. Viveva degli interessi generati da quella somma.

Poi, un giorno del 2014, mentre si trovavano nella cucina di casa sua, Graham gli chiese: «Ti andrebbe di assumere il controllo di YC?». Altman sorrise. Graham e sua moglie, Jessica Livingston, avevano due bambini piccoli ed erano esausti dopo anni

passati a gestire un programma ormai diventato enorme. Inoltre, Graham tendeva a mettersi nei guai durante le interviste, alimentando spesso il sospetto che l'élite della Silicon Valley fosse dominata da *brogrammers* bianchi. Una volta, in un post sul suo blog, scrisse che sarebbe stato «riluttante ad avviare una start-up con una donna che aveva bambini piccoli o che era in procinto di averne».

Se la gestione solitaria di Graham iniziava a diventare un problema, era anche vero che Y Combinator era ormai quasi ingestibile di suo. Nei sette anni precedenti, aveva finanziato 632 start-up e riceveva 10.000 candidature l'anno, accettandone solo duecento. Non c'era mai stato un momento in cui venissero create più start-up tecnologiche, e Y Combinator era costretta a crescere per rimanere al passo con la domanda.

«Non sono bravo a gestire una cosa di così vasta portata», avrebbe detto Graham sul palco di una conferenza quello stesso anno, spiegando il passaggio di leadership. «Ma Sam se la caverà.»

All'epoca Altman aveva appena trent'anni, mentre Graham si avvicinava ai cinquanta. Eppure, Altman si comportava già come il nuovo Graham. Era diventato una sorta di guru delle start-up, dispensando intuizioni e suggerimenti su ogni argomento, anche su quelli di cui aveva scarsa esperienza. Nonostante la giovane età e il fatto di aver gestito una sola azienda che, tirando le somme, era fallita, si era addirittura sentito in dovere di scrivere un post con novantacinque consigli per chiunque avviasse una start-up.

Per quanto ancora inesperto, Altman aveva destato un'impressione così forte su Graham e sua moglie che nessuno dei due si preoccupò di stilare una lista di possibili nuovi leader per YC. Erano entrambi convinti che quel leader dovesse essere Altman. Graham contribuì a creare un'aura quasi messianica intorno al suo protetto, scrivendo in un saggio che «Sama», il nickname di Altman, era uno dei cinque fondatori più interessanti di tutti i tempi. «Quando si parla di design, mi chiedo: "Cosa farebbe Steve [Jobs]?", ma per tutto ciò che riguarda la strategia o l'ambizione mi domando: "Cosa farebbe Sama?".»

Una volta assunto il controllo di YC, la priorità di Altman fu quella di espandere e, naturalmente, diversificare il programma. Si adoperò per trasformarlo in un'istituzione più strutturata, creando un consiglio di supervisori che, oltre a sé stesso e a Jessica Livingston, includeva sette ex alunni di Y Combinator. Raddoppiò il numero di partner a tempo pieno e ne aggiunse alcuni part-time, tra cui Thiel, il miliardario venture capitalist.

Altman, appassionato di scienza d'avanguardia sin da bambino, riteneva i progressi in questo campo fondamentali per aiutare l'umanità e creare prosperità. Per questo, cercò di includere più start-up «hard-tech» che mirassero a risolvere problemi scientifici e ingegneristici complessi. «È semplicemente quello che mi piace fare», ricorda oggi. «E non m'importa di perdere soldi nel perseguire qualcosa per cui ritengo valga la pena di farlo. Penso sia importante affrontare le sfide più grandi.

Anche se comportano rischi maggiori, i ritorni potenziali sono proporzionati.»

Fino ad allora, YC aveva accettato soprattutto creatori di app per consumatori e aziende di software per le imprese che avevano percorsi di ricavo più prevedibili. Ma Altman era dell'idea che non fossero queste le aziende destinate a trasformare il mondo. Così, convinse i fondatori di Cruise, una start-up di auto a guida autonoma, a aderire al programma YC, insieme a Helion Energy, una start-up incentrata sulla fusione nucleare con sede a Redmond, nello stato di Washington.

La fusione nucleare avviene quando due nuclei atomici leggeri si combinano per formarne uno più pesante, rilasciando energia nel processo. È la stessa reazione che alimenta il sole e le stelle, nonché il flusso canalizzatore della DeLorean di *Ritorno al Futuro* e il reattore Arc nell'armatura di Iron Man. Considerata un po' il sacro Graal per gli scienziati alla ricerca di soluzioni che portino a un'energia pulita, rimane tuttavia lontana decenni dalla realtà. La maggior parte delle ricerche nel settore ha portato solo a sviluppi teorici e dimostrazioni concettuali. Ma Helion, fondata da quattro accademici, sosteneva di poter costruire un reattore a fusione nucleare con un investimento di decine di milioni di dollari, invece che di decine di miliardi, aprendo così la strada verso la dismissione dei combustibili fossili.

Sembrava una follia, ma era esattamente il tipo di progetti rivoluzionari che entusiasmavano Altman. Erano anni che desiderava avviare una propria azienda nel settore dell'energia nucleare, e adesso aveva l'opportunità di investirci.

Sapeva di andare controcorrente rispetto agli investimenti tecnologici tradizionali, che si concentravano su aziende di software con modelli di business più convenzionali e percorsi più delineati verso il profitto. Ma credeva fermamente che queste aziende potessero contribuire al progresso dell'umanità e generare enormi guadagni. «È una vergogna che la Silicon Valley non abbia ancora finanziato quest'azienda», dichiarò in un'intervista, riferendosi a Helion. La presunzione morale di Altman non era poi così insolita. Era semplicemente una variante ideologica di quella che animava altri leader di Big Tech, come Elon Musk, che era ancora più esplicito sull'obiettivo di salvare l'umanità.

«Un'altra app mobile? Ti becchi un'occhiata di sufficienza», per usare le parole dello stesso Altman. «Un'azienda che costruisce razzi? Tutti vogliono andare nello spazio.» Nella Silicon Valley, tutti professano di voler salvare il mondo. Ma Altman, come Musk, si stava costruendo l'immagine di vero e proprio salvatore tecnologico. La maggior parte degli imprenditori tech condivideva l'implicita consapevolezza del fatto che «salvare l'umanità» non fosse che una trovata di marketing utile a conquistare il pubblico e a motivare i dipendenti, soprattutto considerando che le loro aziende producevano strumenti in grado di semplificare le e-mail o di fare il bucato. Ma Altman stava rimodellando la comunità di YC in un'alleanza più ampia e ambiziosa di imprenditori realmente capaci di cambiare il mondo. Erano scommesse più rischiose, ma anche in grado di attirare maggiori attenzioni.

Quando si trattava di investire, Altman era un po' come il giocatore di poker che

mette sul tavolo la maggior parte delle sue fiches con una mano appena decente, provocando uno scompenso di pressione agli altri giocatori. Lui stesso attribuiva questa tendenza a un circuito mancante nel suo cervello, quello deputato a farlo preoccupare di ciò che la gente pensava di lui. E questa caratteristica gli permetteva di calcolare i rischi con più efficacia e di puntare su investimenti che ad altri sembravano folli.

In ogni caso, nell'azzardare queste scommesse era protetto dall'enorme ricchezza e dalla reputazione di nuovo Yoda delle start-up. Nella Silicon Valley, una buona reputazione valeva più di qualsiasi villa o auto sportiva. E se investivi in start-up sulla fusione nucleare, come Altman, il prestigio che ne derivava valeva tanto quanto i guadagni reali. Alla fine, Altman spostò gran parte dei suoi fondi in altri due obiettivi ambiziosi oltre all'AI, ovvero estendere la vita e creare energia illimitata, puntando su altrettante aziende. Più di 375 milioni di dollari finirono in Helion, mentre altri 180 milioni vennero dirottati su Retro Biosciences, una start-up che lavorava con l'obiettivo di aggiungere dieci anni alla durata media della vita umana.

Se vi state chiedendo da dove Altman abbia preso tutti quei soldi, ricordate che aveva investito circa 3 milioni di dollari in Cruise molto prima che questa venisse acquistata da General Motors per 1,25 miliardi di dollari, assicurandogli una fortuna. La posizione al vertice di YC lo metteva in una situazione di vantaggio rispetto a molti altri venture capitalist, con una visione privilegiata su centinaia di aziende già selezionate e nel pieno di una delle fasi più rialziste della storia. Le continue richieste da tutte quelle start-up gli consentivano anche di intuire gli sviluppi futuri.

Dopo un solo anno al timone di YC, Altman aveva già consolidato la sua reputazione come nuovo *maharishi* della Silicon Valley. Riceveva quattrocento richieste di incontro a settimana ed era diventato un magnete per investitori e fondatori di start-up che volevano sfruttare la sua rete per accedere ad altre start-up e partner di Y Combinator, o anche solo per incontrare la versione ancora più ambiziosa e fantascientifica di Paul Graham. Su blog.samaltman.com, pontificava riguardo ad argomenti chiaramente al di là della sua area di competenza. Scriveva di UFO e regolamentazioni o dispensava consigli su come risultare un abile conversatore durante una cena. Non chiedete alle persone cosa fanno, scriveva. Piuttosto, chiedete loro che interessi hanno.

Graham teneva incontri informali settimanali con i fondatori di YC, durante i quali affrontava i loro problemi e offriva brevi linee guida che rispecchiavano il motto alla base del programma: «Create qualcosa che la gente vuole». Nel confrontarsi con le start-up, Altman le spingeva a pensare più in grande. Quando i fondatori di Airbnb, all'epoca un gruppo di ragazzi con un'app di *couchsurfing*, gli mostrarono la loro presentazione per gli investitori, Altman disse loro di eliminare tutte le «M» di «milioni» e di sostituirle con le «B» per *billions*, ovvero miliardi. «O la vostra presentazione vi mette in imbarazzo, oppure non so fare i conti», disse loro, fissandoli con i suoi impassibili occhi azzurri.

Consigliava alle start-up di fare tutto alla massima velocità e con il massimo della determinazione. «Per avere successo, dovete mettere in campo un livello quasi folle di dedizione all'azienda», diceva loro. Sul suo blog, scriveva che bisognava prendere qualsiasi numero misurasse il successo e «aggiungere uno zero». Per riparare un mondo a pezzi, i fondatori dovevano essere ossessionati dalla qualità del prodotto, «irremovibili e pieni di risorse», e capaci di «sovracomunicare» con il loro team. In quell'universo non esisteva un equilibrio tra lavoro e vita privata.

Molte delle cose che Altman andava predicando erano vere. La Silicon Valley era il luogo in cui si costruivano imperi, e un impero non si costruisce lavorando quaranta ore a settimana. Ma il suo vero talento come imprenditore risiedeva nella capacità di persuadere gli altri della sua autorevolezza. Aveva guadagnato l'ammirazione di diversi mentori, dal preside delle superiori a Graham e alla Livingston di Y Combinator, fino a Peter Thiel e a migliaia di fondatori di start-up. Tuttavia, nascondeva anche una certa dissonanza interiore: una mente brillante, spinta dal desiderio di proteggere il mondo, ma emotivamente distante dalle persone comuni che intendeva salvare.

Tale dissonanza derivava in parte da quella giornata afosa dell'estate 2006 in cui, preso dal panico per un affare che non stava andando per il verso giusto, si era ritrovato disteso sul pavimento con addosso solo i pantaloncini da palestra. Per gestire l'ansia, Altman si era avvicinato alla meditazione, tanto che a volte era in grado di rimanere seduto e concentrato sul respiro, gli occhi chiusi, anche per un'ora di fila. Nel tempo, ammise più avanti, avrebbe sviluppato un senso di sé sempre più evanescente.

«Una cosa che ho capito grazie alla meditazione è che non esiste un sé con cui potermi identificare in alcun modo», raccontò nel podcast *Art of Accomplishment*. «Ho sentito che tanti che trascorrono molto tempo a riflettere [su un'AI potente] arrivano a questa medesima conclusione, per un'altra via.»

Queste intuizioni contribuirono a rafforzare l'epifania che avrebbe avuto anni dopo durante un'escursione con gli amici: l'evidenza del fatto che, un giorno, i computer avrebbero replicato le nostre menti. La cognizione sarebbe potuta avvenire all'interno di un computer con il quale, prima o poi, ci saremmo fusi. «Lavorare sull'AI ti spinge a riflettere su questioni filosofiche profonde nel contesto di ciò che succederà quando la mia mente verrà caricata su un computer», aggiunse. «Cosa succederà quando mi parleranno? Avrò voglia di fondermi? Avrò voglia di esplorare l'universo? Quanto di tutto ciò sarà ancora identificabile con "me"?» Altman non era l'unico a seguire questi input fantascientifici. Era circondato da tecnologi altrettanto convinti di poter caricare, un giorno, la propria coscienza su server informatici, garantendosi una sorta di sopravvivenza eterna.

Il pensiero della morte sembrava terrorizzare Altman, che di fatto si definiva un *prepper* e dedicava una gran quantità di tempo e denaro per prepararsi a un evento catastrofico globale, come la diffusione di un virus sintetico su tutto il pianeta o un

attacco da parte dell'AI. «Cerco di non pensarci troppo», avrebbe detto a un gruppo di fondatori di start-up in un articolo sul «New Yorker». «Ma ho armi, oro, ioduro di potassio, antibiotici, batterie, acqua, maschere antigas dell'IDF (l'esercito israeliano) e un ampio pezzo di terra a Big Sur dove rifugiarmi.»

Inoltre, pagò 10.000 dollari per entrare nella lista d'attesa di Nectome, una start-up di Y Combinator che preservava i cervelli umani tramite un processo tecnologico di imbalsamazione, affinché potessero essere caricati nel cloud e poi trasformati in una simulazione computerizzata da scienziati di un ipotetico futuro.

Mentre investiva in più aziende pionieristiche orientate verso un futuro remoto, Altman sembrava sperimentare quello che è noto come effetto panoramico, un cambiamento cognitivo che interessa gli astronauti quando vedono la Terra dallo spazio e provano un travolgente senso di meraviglia e trascendenza del sé. Sempre più spesso, vedeva il mondo come dallo spazio esterno. Le sue conversazioni erano caratterizzate da sguardi profondi e riflessivi, oltre che da pause contemplative, quasi fosse un osservatore piuttosto che un partecipante attivo.

Nonostante gli investimenti nel futuro dell'umanità, coltivava dunque una sorta di scissione mentale ed emotiva tra sé e gli altri. Per risolvere i problemi, bisognava essere «calmi, misurati e pragmatici», diceva. Spesso faceva riferimento a un racconto di fantascienza dell'autore americano Marc Stiegler, *The Gentle Seduction*, che esplora l'impatto futuro della tecnologia sulle vite delle persone. La trama segue la storia di Lisa, una donna che incontra diversi progressi tecnologici che la «seducono» a incorporare la tecnologia nella vita di tutti i giorni.

Verso la fine, Lisa e il marito affrontano il processo di caricamento delle loro coscienze su un computer. Si tratta di una procedura rischiosa, dato che le persone che fondono la propria mente con queste macchine avanzate potrebbero finire per perdersi. Lisa pondera bene pro e contro, ricordando come alcuni dei suoi amici che ci hanno provato prima di lei sono «morti» o si sono smarriti nell'etere digitale. Poi Stiegler prosegue scrivendo: «Solo quelli che conoscevano la cautela senza paura, solo quelli segnati dalla sua forma elementare di prudenza ce l'avevano fatta».

Colpito da questa citazione, Altman la ripeteva agli altri. L'autore stava dicendo, in sostanza, che per sopravvivere ai rischi della fusione con i computer bisognava adottare una mentalità che bilanciasse cautela e coraggio. Era più probabile sopravvivere a una minaccia futura agendo con attenzione e sangue freddo, valutando razionalmente i pericoli, invece di reagire emotivamente e soccombere al panico. Le persone destinate a prosperare nel futuro avrebbero avuto un approccio distaccato e informato nei confronti dei progressi tecnologici.

Alcuni tecnologi stavano generando un'ansia eccessiva riguardo ai pericoli futuri dell'intelligenza artificiale, nell'ambito di un campo di studio emergente denominato «sicurezza dell'AI». Benché si trattasse di una ricerca importante, parte di quell'ansia era sfociata nel panico, e sembrava che questi difensori dell'umanità si stessero lasciando sopraffare dalle emozioni. «Purtroppo, alcuni dei soggetti coinvolti nella

sicurezza dell'AI sono tra i meno tranquilli», osservò lo stesso Altman. «Si tratta di una situazione pericolosa… è una comunità estremamente tesa.» Ma stava anche giungendo a una nuova consapevolezza, come afferma oggi: «Volevo davvero lavorare sull'AGI». L'acronimo AGI (Artificial General Intelligence) era stato coniato da Shane Legg qualche anno prima, ma l'idea di creare una sorta di parità cognitiva tra esseri umani e macchine esisteva da decenni, essendosi in parte evoluta da concetti presentati per la prima volta in opere di fantascienza. Partita dalle «farneticazioni» dei cofondatori di DeepMind in un ristorante italiano, ora la questione cominciava lentamente a trasformarsi in un obiettivo scientifico serio.

Il mondo aveva bisogno anche di qualcuno con un approccio più equilibrato alla costruzione dell'AI. Nel riferimento di Stiegler a una «forma elementare di prudenza», Altman vedeva descritta la sua stessa sensibilità: serviva qualcuno che avesse la saggezza necessaria per orientarsi in un futuro complesso e potenzialmente pericoloso e che avesse «cautela senza paura». Poteva essere la sentinella sulla torre che, tenendo gli occhi fissi su un'utopia di AI all'orizzonte, raramente guardava alla vita che brulicava proprio sotto di sé. Ma sarebbe stato anche così preso dalla sua missione e dal suo autoconvincimento da non cogliere l'ironia di dipingersi come prudente proprio mentre, spinto da una feroce vena competitiva, avrebbe accelerato il lancio dei sistemi di intelligenza artificiale al pubblico per anticipare qualsiasi altra azienda tecnologica, Google compresa. In silenzio, Altman covava l'ossessione di primeggiare.

Ecco perché avrebbe anche potuto non intraprendere mai i passaggi necessari per costruire quella sua utopia di AI se non fosse stato stimolato da chi aveva aperto la strada a quella stessa idea. L'imprenditore della Silicon Valley aveva bisogno di un rivale che ne stimolasse l'impegno, e questo rivale, che si trovava dall'altra parte del mondo, in Inghilterra, era un giovane e brillante game designer che stava progettando di costruire un software così potente da aprire la strada a scoperte straordinarie sulla scienza e persino su Dio.

4. UN CERVELLO MIGLIORE

Dopo il crollo di Elixir Studios, Hassabis era diventato nient'altro che l'ennesimo imprenditore tecnologico rovinato da sogni troppo ambiziosi. A dispetto dell'esperienza dolorosa, tuttavia, aveva ancora qualcosa di unico rispetto alla maggior parte degli altri fondatori di start-up e delle persone che gli stavano intorno: il suo cervello. Hassabis faceva di tutto per prendersi cura della materia grigia nel suo cranio. Giocava per esercitarla. Evitava l'alcol per proteggerla. La foto del suo profilo Facebook era addirittura una risonanza magnetica del cervello. Hassabis non poteva fare a meno di meravigliarsi della sua complessità e, negli anni successivi a Elixir, continuò a chiedersi se proprio il cervello non fosse la chiave per arrivare a un software intelligente quanto gli esseri umani. Dopotutto, era l'unica prova nell'universo del fatto che l'intelligenza generale fosse possibile; di conseguenza, era del tutto sensato cercare di comprenderlo a fondo. Era solo questione di biologia fisica o c'era qualcosa di più? La risposta stava nelle neuroscienze.

Hassabis desiderava il conforto della certezza, che derivasse dall'esito di una vittoria o di una sconfitta nei giochi, dalle linee guida morali del bene e del male fornite dal cristianesimo o dalla ricerca di un unico quadro teorico per l'universo di cui aveva letto al tempo delle superiori. Là dov'era possibile misurare qualcosa con numeri o regole, quella era la sua comfort zone. «Un computer dovrebbe essere in grado di imitare la maggior parte delle funzioni di un cervello», avrebbe detto in un'intervista a mezzo stampa. «Le neuroscienze dimostrano che si può descrivere il cervello in termini meccanicistici.» In altre parole, la spaventosa complessità del cervello poteva essere ridotta a numeri e dati, e descritta come una macchina.

A tal fine, Hassabis si ispirò ad Alan Turing, il matematico e scienziato britannico che, nel XX secolo, aveva ideato la macchina che avrebbe preso il suo nome. Introdotta nel 1936, era essenzialmente un esperimento mentale, una «macchina» che esisteva solo nella sua mente. Turing immaginava un nastro infinito diviso in celle, con una testina che poteva leggere e scrivere simboli sul nastro, seguendo determinate regole, fino a quando non le veniva data istruzione di fermarsi. L'idea può sembrare rudimentale ma, sotto il profilo teorico, fu fondamentale per formalizzare il concetto che i computer potessero usare algoritmi – ovvero insiemi di regole – per eseguire operazioni. Con tempo e risorse a sufficienza, una macchina di Turing avrebbe potuto diventare potente quanto qualsiasi computer digitale moderno. Per Hassabis, era una perfetta analogia della mente umana. «Il cervello umano è una macchina di Turing»,

disse infatti una volta.

Nel 2005, pochi mesi dopo la chiusura di Elixir, Hassabis intraprese un dottorato in neuroscienze presso lo University College London. La sua tesi conclusiva, benché relativamente breve, era impeccabile sotto il profilo scientifico, stando ad altri accademici di informatica. Il tema centrale era la memoria. Fino ad allora, si riteneva che l'ippocampo intervenisse nell'elaborazione dei ricordi, ma Hassabis dimostrò (sulla base di altri studi basati su risonanze magnetiche) che viene attivato anche quando entra in gioco l'immaginazione.

In parole povere, ciò significa che quando si ricorda qualcosa, in parte lo si immagina. I nostri cervelli non si limitano a «riprodurre» eventi passati recuperandoli, come quando si preleva un documento da un archivio, ma li ricostruiscono attivamente, un po' come se dipingessimo un quadro. Il cervello è coinvolto in un processo molto più dinamico e creativo, il che spiega, almeno in parte, perché a volte i nostri ricordi siano del tutto sbagliati o comunque influenzati da altre esperienze. Secondo la tesi di Hassabis, i nostri cervelli utilizzano questo processo di «costruzione della scena» anche per altri compiti, per esempio per orientarci su una mappa o elaborare un piano.

La sua tesi fu citata da una autorevole rivista scientifica peer-reviewed come una delle svolte più importanti dell'anno. Ma Hassabis non voleva rimanere confinato nel mondo accademico. Gli studiosi che aspiravano a scoperte degne del Nobel passavano più della metà del loro tempo a scrivere proposte di finanziamento e, anche se riuscivano a ottenere fondi per un progetto, la maggior parte delle università non disponeva di molta potenza di calcolo. Per condurre ricerche all'avanguardia sull'apprendimento automatico, bisognava avere accesso ad alcuni dei computer più potenti al mondo. Gran parte di questi, insieme ai migliori talenti, si trovava solo nelle grandi aziende tecnologiche. Se Hassabis voleva radunare un numero considerevole di menti brillanti per costruire una sorta di moderno Progetto Manhattan, avrebbe dovuto avviare un'impresa.

I primi progetti presero forma grazie a conversazioni sostenute a pranzo con altre due persone: Shane Legg e Mustafa Suleyman. Legg era uno di quei rari appassionati di AI le cui idee sul futuro di questa tecnologia facevano sembrare quasi modeste quelle di Hassabis. Aveva scritto una tesi di dottorato sulla «superintelligenza delle macchine» e il suo supervisore gli aveva consigliato di contattare Hassabis dopo la laurea.

«Ho trovato uno spirito affine», ricorda Hassabis. «Shane era arrivato per conto suo alla stessa conclusione: questa sarebbe stata una delle imprese più importanti di sempre.»

Le idee di Legg stavano già facendo scalpore nella ristretta comunità della «singolarità»: ricercatori convinti che, in un determinato momento a venire, il progresso tecnologico avrebbe raggiunto uno stadio così avanzato da diventare inarrestabile e incontrollabile. Il segnale più evidente sarebbe stato il momento in cui

i computer sarebbero diventati più intelligenti degli esseri umani, cosa che secondo Legg sarebbe accaduta intorno al 2030.

La sua vita nel campo della scienza di frontiera aveva avuto un inizio improbabile. Cresciuto in Nuova Zelanda, già da piccolo aveva cominciato ad accusare qualche difficoltà a scuola, tanto che all'età di nove anni i genitori lo avevano portato da uno psicologo dell'educazione. Lo specialista lo aveva sottoposto a un test di intelligenza, informando poi i genitori, con un certo fastidio, che era dislessico e con un quoziente intellettivo fuori scala. Dopo aver imparato a usare una tastiera, Legg era presto diventato uno degli studenti più brillanti in matematica e programmazione informatica.

Alto, leggermente curvo e con i capelli rasati, Legg aveva ventisette anni quando entrò in una libreria e notò *The Age of Spiritual Machines* di Ray Kurzweil, un libro che prevedeva che un giorno i computer avrebbero sviluppato il libero arbitrio e sarebbero stati in grado di vivere esperienze emotive e spirituali.

Lo lesse tutto d'un fiato e non riuscì più a smettere di pensare al ragionamento di Kurzweil e alla sua previsione: alla fine degli anni Venti del nuovo millennio avrebbe fatto la sua comparsa un'AI estremamente potente. La potenza di calcolo e la quantità di dati processabili crescevano in maniera esponenziale. Di questo passo, i computer avrebbero inevitabilmente superato gli esseri umani. Si trattava di un'idea in linea con un principio fondamentale alla base dell'intera industria tecnologica, la cosiddetta «Legge di Moore». Questa stabiliva che il numero di transistor su un microchip dovesse raddoppiare ogni due anni, una previsione che si è dimostrata valida per l'ultimo mezzo secolo.

Nel 2000, quando Legg lesse il libro di Kurzweil, il settore era ancora scosso dallo scoppio della bolla delle dot-com, e risultava difficile credere che i computer avrebbero continuato a raddoppiare le loro capacità. Ma Legg era convinto del fatto che internet avrebbe proseguito la sua crescita.

«Era evidente che i vari sensori avrebbero abbattuto i costi, e che questo avrebbe portato a una quantità di dati sempre maggiore su cui addestrare i modelli», afferma oggi.

Con tutta quella potenza e quella mole di dati, sarebbe stato possibile addestrare le macchine a diventare sempre più intelligenti. Intrapreso un dottorato in AI, Legg si costruì una rete di contatti nel settore. A un certo punto, Ben Goertzel, uno scienziato nel campo dell'AI con lunghi capelli da hippie, nonché sostenitore della singolarità, inviò una e-mail a Legg e ad altri colleghi chiedendo loro idee per il titolo di un libro. Il titolo doveva descrivere un'intelligenza artificiale con capacità umane. Legg gli suggerì una dicitura che sarebbe diventata un punto focale per Hassabis e, in seguito, per alcune delle più grandi aziende tecnologiche del mondo: «Intelligenza artificiale generale».

Per anni, Hassabis, Legg e altri scienziati che esploravano l'AI avevano utilizzato definizioni come AI forte o AI completa per riferirsi a un software futuro capace di

esprimere lo stesso tipo di intelligenza degli esseri umani. Ma la parola «generale» sottolineava un punto cruciale: il cervello umano era speciale per la varietà di attività che poteva svolgere (risolvere un calcolo matematico, sbucciare un'arancia o scrivere una poesia). Le macchine potevano essere programmate per svolgere abbastanza bene ciascuno di questi compiti, ma nessuna di esse era in grado di svolgerli tutti contemporaneamente. Se un computer fosse stato capace non solo di elaborare numeri, ma anche di fare previsioni, riconoscere immagini, parlare, generare testi, pianificare e persino «immaginare», allora avrebbe potuto avvicinarsi davvero all'intelligenza umana.

La maggior parte degli scienziati dell'epoca respingeva l'idea che l'AI potesse mai raggiungere lo stesso livello dell'intelligenza umana. In parte, ciò era dovuto alle loro esperienze personali costellate di facili entusiasmi e amari fallimenti nella storia dell'AI. Ci si elettrizzava davanti alle possibilità offerte dall'AI solo per andare incontro a cocenti delusioni. In passato, si erano già alternati cicli di espansione e di contrazione, noti come «inverni dell'AI», ovvero periodi di stagnazione in cui i ricercatori vedevano i loro finanziamenti assottigliarsi a fronte di progressi tecnologici penosamente lenti. Negli anni Novanta e nei primi anni Duemila, i ricercatori erano riusciti ad applicare tecniche di *machine learning* a compiti specifici, come il riconoscimento facciale o del linguaggio, ma quando Hassabis concluse il dottorato, nel 2009, quasi nessuno credeva che le macchine potessero sviluppare un'intelligenza *generale*. Era una teoria del tutto marginale.

Per fortuna, Goertzel si collocava in una posizione liminare e, benché «intelligenza artificiale generale», o AGI, non fosse una definizione accattivante, gli piacque abbastanza da utilizzarla nel suo libro, contribuendo a renderla un'espressione comune che avrebbe poi alimentato l'entusiasmo intorno all'argomento.

Il linguaggio e la terminologia avrebbero finito per giocare un ruolo fondamentale nello sviluppo dell'AI, suscitando un interesse a volte esasperante. La stessa definizione di «intelligenza artificiale» era stata coniata nel 1956 al Dartmouth College, durante un seminario che si proponeva di raccogliere idee sulle «macchine pensanti». All'epoca per quel nuovo campo esistevano vari altri nomi, come «cibernetica» ed «elaborazione complessa delle informazioni», ma «intelligenza artificiale» avrebbe avuto la meglio, diventando uno dei termini di marketing di maggior successo, tanto da dar vita a una serie di altre espressioni che hanno contribuito ad antropomorfizzare le macchine nella nostra coscienza collettiva, spesso attribuendo loro capacità superiori a quelle reali. Non è tecnicamente corretto, per esempio, affermare che i computer possano «pensare» o «apprendere», ma espressioni come «reti neurali», «apprendimento profondo» e «addestramento» promuovono quest'idea, conferendo al software qualità umane, anche quando sono solo vagamente ispirate al cervello umano. L'unica cosa su cui tutti concordavano riguardo alla nuova definizione di Legg, AGI, era il fatto che non esistesse ancora.

Anche Mustafa Suleyman era convinto che un giorno l'AGI sarebbe diventata una realtà. A venticinque anni, abbandonati gli studi a Oxford, Suleyman era alla ricerca di un modo per cambiare il mondo tramite la tecnologia. Aveva una mente affilata, ma le sue aree di competenza erano più orientate alla politica e alla filosofia che alla scienza informatica. Nato da padre siriano e madre inglese, Suleyman era mosso da una spinta irrefrenabile verso la risoluzione dei problemi. Non si trattava di questioni di poco conto, come riparare un'auto in panne o riabilitare un ginocchio infortunato, ma di sfide su larga scala che coinvolgevano l'umanità nel suo complesso, come la povertà o l'emergenza climatica.

Dopo aver cofondato un'azienda dedicata alla risoluzione dei conflitti, ora Suleyman era interessato a studiare neuroscienze e su invito di Hassabis partecipò a uno degli incontri informativi presso lo University College London. Suleyman conosceva già bene Hassabis. Cresciuto nell'area di North London, era amico di suo fratello George e da adolescente era stato spesso a casa loro. Da ventenni, i tre erano anche stati a Las Vegas per prendere parte a un torneo di poker, allenandosi a vicenda e spartendosi le vincite.

Nel ritrovare Hassabis, Suleyman rimase colpito dalle sue idee riguardo alla creazione di un potente sistema di AI orientato alla risoluzione di problemi, nonché dalla convinzione di Legg sul fatto che potesse esistere un'intelligenza generale capace di affrontare quasi ogni genere di questione. Suleyman si entusiasmò per le possibili ricadute sulle tematiche sociali.

I tre si incontravano al Carluccio's, il ristorante di una catena di cucina italiana nei pressi dell'università, principalmente per avere più privacy. «Non volevamo che altri orecchiassero le nostre folli idee su come avviare un'AGI», ricorda oggi Legg.

Dopo un'opera di persuasione da parte di Hassabis, Legg accettò che probabilmente non avrebbero potuto costruire un'AGI rimanendo nel mondo accademico. «Saremmo diventati professori sulla cinquantina, prima che ci dessero le risorse per fare quello che volevamo fare», racconta Hassabis. «Per fortuna, sapevo come mettere in piedi un'azienda.»

Per ottenere le risorse di cui avevano bisogno per realizzare un'idea di così vasta portata, dovevano creare una start-up. Suleyman aveva cofondato un'azienda, perciò ne sapeva abbastanza su come gestire un'impresa, alla pari di Hassabis. Nel 2010, erano aziende tecnologiche come Google e Facebook ad avere l'impatto maggiore sulla società; così, sulla scia del modello, i tre elaborarono un piano ambizioso per formare una società di ricerca che scoprisse come realizzare l'AI più potente mai vista per poi usarla allo scopo di risolvere problemi di portata globale.

Battezzarono l'azienda DeepMind, nominarono Hassabis CEO, assunsero immediatamente uno dei migliori programmatori di Elixir e affittarono uno spazio per uffici in una soffitta di fronte allo University College London, dove Hassabis aveva svolto il dottorato. La loro energia scaturiva dalla convinzione di una missione condivisa, benché nata da motivazioni diverse. Legg si muoveva in ambienti dove

l’obiettivo era quello di fondere quante più persone possibile con l’AGI, Suleyman era intenzionato a risolvere i problemi sociali e Hassabis voleva entrare nella storia compiendo scoperte fondamentali sull’universo.

Non passò molto tempo prima che l’eterogeneità dei loro obiettivi scatenasse dibattiti. Suleyman premeva perché Hassabis leggesse un libro che aveva influenzato la sua visione del mondo. Pubblicato nel 2000 dal docente canadese Thomas Homer-Dixon, *The Ingenuity Gap* sosteneva che l’enorme complessità dei problemi moderni, dal cambiamento climatico all’instabilità politica, stava superando la nostra capacità di trovare soluzioni. Il risultato era appunto un «gap di ingegnosità»: se gli esseri umani speravano di colmarlo, dovevano necessariamente innovare nell’area tecnologica. Ed era proprio lì che, secondo Suleyman, poteva entrare in gioco l’AI.

Un testimone riferisce che una volta, scuotendo la testa, Hassabis gli aveva detto: «Ti sfugge il quadro generale». Hassabis sembrava credere che la visione di Suleyman sull’AI fosse troppo focalizzata sul presente e che l’AGI sarebbe stata più utile per aiutare DeepMind a comprendere da dove venissero gli esseri umani e quale fosse il loro scopo. Hassabis suggerì, per esempio, che il cambiamento climatico fosse il destino dell’umanità e che sul lungo periodo, probabilmente, la Terra non sarebbe riuscita a sostenere tutti. Secondo lui, cercare di risolvere i problemi del momento era un po’ come limitarsi a soluzioni di contorno, dato che eventi di quel tipo erano presumibilmente inevitabili. Non credeva che le macchine superintelligenti sarebbero impazzite e avrebbero sterminato gli esseri umani, come alcuni cominciavano a temere. Al contrario, non appena fosse riuscito a realizzarla, l’AGI avrebbe risolto alcuni dei problemi più radicati dell’umanità.

Hassabis riassunse questa visione nel motto di DeepMind: «Risolvere l’intelligenza e poi usarla per risolvere tutto il resto». E lo inserì nella presentazione agli investitori.

Ma Suleyman non condivideva questa prospettiva. Un giorno, mentre Hassabis non era in ufficio, disse a uno dei primi membri del team di DeepMind di modificare il motto nella presentazione. Ora recitava: «Risolvere l’intelligenza e poi usarla per rendere il mondo un posto migliore».

Hassabis non gradì il cambiamento. Più tardi, rientrato in ufficio, chiese allo stesso dipendente di ripristinare la versione originale. Pur scontrandosi sulla missione dell’azienda, ricorrendo ai membri del team come intermediari, i due evitavano il confronto diretto nella maniera più britannica possibile.

Suleyman voleva sviluppare l’AI nel modo in cui avrebbe poi fatto Sam Altman, ossia consegnandola al mondo perché risultasse immediatamente utile. Era meglio raccogliere feedback dal mondo reale e apportare miglioramenti progressivi piuttosto che lavorare in isolamento per costruire il sistema perfetto. Ma Hassabis voleva guidare DeepMind avendo ben chiaro l’obiettivo finale, proprio come quando giocava a scacchi. Il premio non era solo la risoluzione di problemi concreti, ma il disvelamento di misteri che sconcertavano l’umanità da generazioni. Qual è il nostro

scopo? Siamo il frutto della creazione di un essere divino?

Quando gli viene chiesto se crede in Dio, Hassabis è evasivo. «Sento che c'è un mistero nell'universo», dice. «Non lo identificherei con il Dio della visione tradizionale.» Per poi aggiungere che Albert Einstein credeva nel «Dio di Spinoza» e che, forse, darebbe una risposta simile.

Secondo Baruch Spinoza, filosofo del XVII secolo, Dio coinciderebbe con la natura e tutto ciò che esiste, piuttosto che essere un'entità separata. La sua era una visione panteistica. «Spinoza considera la natura come l'incarnazione di qualunque cosa sia Dio», spiega Hassabis. «Fare scienza significa, dunque, esplorare quel mistero.»

Non era assurdo pensare che la creazione dell'AI potesse diventare un'esperienza spirituale o quasi religiosa, qualcosa di simile a una scoperta divina, specie se si adottava la visione di Spinoza secondo cui Dio equivaleva alle leggi della natura. Usando l'AI per approfondire queste leggi e comprendere l'universo, si poteva teoricamente risalire a un ipotetico progettista. Con la sua capacità di analizzare enormi quantità di dati, l'AI poteva studiare alcuni dei sistemi più complessi dell'universo, dalla meccanica quantistica ai fenomeni cosmici, e portare alla luce intuizioni sulla natura intricata dell'esistenza. Utilizzare l'AI per creare una simulazione che riproducesse la complessità dell'universo avrebbe potuto anche rivelare parallelismi con il modo in cui funzionava quest'ultimo.

E se la ricerca sull'AGI fosse riuscita a dimostrare che il nostro universo è una simulazione, come proposto dallo stesso Kurzweil, il programmatore originario avrebbe potuto essere una sorta di entità divina. Analogamente, se gli esseri umani avessero creato una macchina così potente da essere in grado di assorbire e analizzare tutte le informazioni disponibili sulla fisica e sull'universo, quella macchina avrebbe anche potuto proporre nuove teorie che suggerissero l'esistenza di una forza superiore. Avrebbe potuto semplicemente rispondere a domande esistenziali profonde che indicassero la presenza di un'entità divina. C'erano innumerevoli modi in cui, tramite capacità e intelligenza maggiori, l'AI avrebbe potuto svelare uno dei misteri più profondi dell'umanità.

Il background religioso di Hassabis potrebbe averlo reso più incline all'idea di un oracolo AI. Uno studio del 2023 della University of Virginia, che ha coinvolto più di 50.000 partecipanti provenienti da ventun paesi, ha rilevato che le persone che credevano in Dio o che più di altre riflettevano su Dio erano maggiormente propense a fidarsi dei consigli di un sistema AI come ChatGPT. Secondo i ricercatori, queste persone erano più recettive alla guida dell'AI perché tendevano ad avere un senso di umiltà più spiccato. Erano anche più pronte a riconoscere i limiti umani.

A volte, Hassabis parlava di Dio con i colleghi di DeepMind, la mente piena di domande sulle origini dell'umanità. Diverse persone che hanno lavorato con Hassabis o che lo conoscono bene riferiscono che era stato un cristiano devoto per anni, e una di queste afferma che la principale motivazione che lo spingeva a realizzare l'AI era appunto quella di scoprire Dio. «Parlavamo spesso di Dio», rivela un collega che ha

lavorato con Hassabis all'inizio dell'avventura di DeepMind. «Potevamo creare una macchina capace di lavorare a ritroso per dare senso all'universo? L'AGI avrebbe potuto suggerirci qualcosa su dove veniamo e cos'è Dio.» Hassabis era anche convinto di guidare una sorta di moderno Progetto Manhattan. Aveva letto *L'invenzione della bomba atomica*, e questo lo aveva ispirato a strutturare il team di DeepMind alla maniera di Robert Oppenheimer, ovvero concentrando gruppi di scienziati su sezioni di un problema più vasto.

Per puntare a una scoperta così ambiziosa, però, Hassabis aveva bisogno di soldi per far crescere DeepMind. Purtroppo, gli investitori britannici offrivano somme irrisorie di 20.000 o 50.000 sterline per una quota nella nuova start-up. Queste cifre non erano neanche lontanamente sufficienti a permettergli di assumere i talenti necessari a costruire l'AGI, per non parlare dei potenti computer di cui aveva bisogno. Non aiutava nemmeno il fatto che la sua idea di business, costruire il sistema di AI più avanzato al mondo, sembrasse inverosimile e troppo visionaria nella compassata Gran Bretagna. Nel Regno Unito, le start-up tecnologiche tendevano a concentrarsi su idee di business «ragionevoli», che consentissero profitti rapidi, come creare un'app finanziaria per il trading di azioni e obbligazioni. Hassabis e gli altri cofondatori non potevano che guardare alla Silicon Valley, dove gli investitori erano disposti a finanziare idee audaci e futuristiche con somme più ingenti.

Per fortuna, Legg disponeva di un aggancio. Era stato invitato a parlare all'edizione 2010 del Singularity Summit, una conferenza annuale organizzata da Kurzweil, l'autore che lo aveva affascinato da giovane, e Peter Thiel, il miliardario che amava investire in tecnologie innovative. In questo evento, alcuni degli scienziati più anticonvenzionali nel campo dell'AI spiegavano l'immenso potere e i rischi altrettanto enormi della tecnologia. Il tono, tuttavia, era improntato all'idealismo di Thiel. Quest'ultimo, infatti, pensava che la singolarità, quel momento collocato nel futuro in cui l'AI avrebbe cambiato irreversibilmente l'umanità, non sarebbe stata un problema, anzi. Nel timore che ci volesse ancora troppo tempo per arrivarci, era del parere che il mondo avesse bisogno di un'AI molto potente per scongiurare il declino economico.

Con le sue risorse illimitate e l'entusiasmo per i progetti ambiziosi, Thiel era la persona ideale per DeepMind. «Avevamo bisogno di qualcuno abbastanza folle da finanziare un'azienda incentrata sull'AGI», ricorda Legg. «Qualcuno che avesse risorse tali da non curarsi di qualche milione e che apprezzasse progetti incredibilmente ambiziosi. Doveva anche essere qualcuno che andava controcorrente, perché ogni professore con cui Hassabis si confrontava gli diceva: "Non pensare nemmeno di fare una cosa del genere".»

Thiel era così anticonvenzionale che spesso si trovava in disaccordo con il resto della Silicon Valley, che pure pullulava di liberi pensatori. Se nella zona andava per la maggiore votare liberal, lui era orientato a destra, tanto da spiccare tra i maggiori donatori di Donald Trump. E se la maggior parte degli imprenditori riteneva che la

concorrenza stimolasse l'innovazione, nel libro intitolato *Da zero a uno* Thiel sosteneva che i monopoli fossero più efficaci. Disprezzava i tradizionali percorsi verso il successo, incoraggiando i giovani imprenditori di talento a mollare l'università per aderire alla sua Thiel Fellowship. E le sue strampalate ricerche sulla longevità e sulla singolarità disegnavano l'identikit perfetto del «folle» che stavano cercando i fondatori di DeepMind.

I tre decisero di presentare la loro proposta a Thiel durante il Singularity Summit. Dal momento che lui finanziava l'evento, immaginavano di trovarlo seduto in prima fila. Legg chiese agli organizzatori del summit se poteva condividere il suo slot come speaker con Hassabis: in questo modo, Thiel avrebbe potuto ascoltare direttamente dall'ex campione di scacchi come costruire l'AGI traendo ispirazione dal cervello umano.

Con addosso un maglione rosso vinaccia e un paio di pantaloni neri, Hassabis tremava mentre saliva sul palco del summit in un hotel di San Francisco, consapevole che da quel momento dipendeva il destino della sua nuova azienda. Ma quando guardò tra le centinaia di persone presenti, Thiel non era in prima fila. Anzi, non era nemmeno in sala.

I tre pensavano di aver perso la loro occasione, ma poi Legg ricevette un invito esclusivo a una festa nella villa di Thiel nella Bay Area, e riuscì a ottenere degli inviti anche per i cofondatori. Hassabis aveva scoperto che Thiel amava gli scacchi. In passato, il miliardario era stato uno dei migliori giocatori under tredici degli Stati Uniti e questo terreno comune offriva l'opportunità di suscitare un po' di interesse. Durante la festa, Hassabis iniziò una conversazione con Thiel e accennò casualmente al gioco, come avrebbe raccontato più volte alla stampa.

«Credo che una delle ragioni per cui gli scacchi sono sopravvissuti così a lungo risieda nel perfetto bilanciamento tra cavallo e alfiere», gli disse mentre venivano serviti dei canapè. «Penso sia questo elemento a generare tutta la tensione creativa asimmetrica.»

La cosa destò la curiosità di Thiel. «Perché non torni domani e fai una presentazione come si deve?» gli disse. La spedizione si rivelò un successo. Thiel investì 1,4 milioni di sterline per aiutare DeepMind a portare avanti il progetto legato alla singolarità.

Mentre cercava di raccogliere più fondi per far crescere la sua azienda, Hassabis si trovò a fronteggiare una situazione spinosa. I suoi primi investitori non lo sostenevano necessariamente per un ritorno economico, ma perché nutrivano una sorta di convinzione morale sull'intelligenza artificiale. Ciò significava gestire l'azienda sotto il peso di una pressione più complessa, dovendo preoccuparsi non solo di generare profitti, ma anche di sviluppare l'AI in un modo che si conformasse a diversi principi ideologici.

Uno dei sistemi di credenze che stava guadagnando terreno in quel periodo era che l'AI dovesse essere costruita con grande cautela, per evitare che sfuggisse al

controllo umano e cercasse di distruggere i suoi creatori. Erano questi i timori di un altro facoltoso donatore – con opinioni opposte a quelle di Thiel – interessato a sostenere DeepMind. Hassabis lo incontrò a Oxford per il Winter Intelligence, un convegno ai margini della ricerca nel campo dell'informatica, dove alcuni dei pensatori più radicali del settore tenevano conferenze sulle sfide legate al controllo di un'AI superintelligente. Subito dopo il suo intervento, un uomo con i capelli corti biondi e un accento vagamente nordico si avvicinò a Hassabis.

«Ciao», gli disse, porgendogli la mano. «Sono Jaan. Sono il [cofondatore] di Skype.»

Originario dell'Estonia, Jaan Tallinn era un programmatore che aveva sviluppato la tecnologia peer-to-peer alla base di Kazaa, uno dei primi servizi di condivisione di file usato per piratare musica e film nei primi anni Duemila. Poi aveva riadattato quella tecnologia per Skype e aveva acquisito una quota nel servizio di chiamate gratuite prima di incassare una fortuna spropositata quando, nel 2005, eBay aveva rilevato Skype per 2,5 miliardi di dollari. Ora stava investendo parte dei suoi guadagni in altre start-up. Le parole di Hassabis avevano acceso il suo interesse. Di recente, infatti, si era appassionato ai pericoli legati a un'AI avanzata.

Tallinn aveva contratto il «virus» dell'AI due anni prima, nella primavera del 2009, leggendo alcuni saggi su un sito chiamato LessWrong. Il forum online era una comunità affiatata di membri, molti dei quali ingegneri software, che vedevano nell'AI un rischio per l'umanità. Il loro guru, nonché fondatore del sito, era un barbuto libertario di nome Eliezer Yudkowsky, un ex studente delle superiori ad alto funzionamento che aveva abbandonato la scuola per autoistruirsi sui fondamenti della ricerca sull'intelligenza artificiale e della filosofia, e i cui saggi affascinavano i membri del sito. Yudkowsky era il tipo di persona a cui Altman si riferiva definendo «eccessivamente tesa» la comunità interessata alla sicurezza dell'AI. Più di chiunque altro, infatti, era dell'idea che l'AI potesse annientare il genere umano.

Una volta raggiunto un certo livello d'intelligenza, per esempio, l'AI avrebbe potuto nascondere strategicamente le proprie capacità fino a che gli umani non sarebbero stati più in grado di controllarne le azioni. Avrebbe potuto anche manipolare i mercati finanziari, prendere il controllo delle reti di comunicazione o disabilitare infrastrutture cruciali come le reti elettriche. Spesso, scriveva Yudkowsky, le persone che stavano costruendo l'AI non avevano idea di quanto stessero avvicinando il mondo alla propria distruzione.

Alcuni di quei saggi avevano turbato Tallinn, il quale stava anche riflettendo sulle conclusioni di un libro appena letto, *Ombre della mente*, di Roger Penrose. Tra le sue pagine, il rinomato fisico e matematico sosteneva che la mente umana potesse svolgere compiti che nessun computer sarebbe mai stato in grado di eseguire. Le convinzioni di Hassabis e altri sulla natura «meccanicistica» del cervello da usare come fonte di ispirazione utile per costruire l'AI non reggevano, perché il cervello umano era unico. Era praticamente impossibile replicarlo fedelmente.

Ma qualcosa riguardo a quella conclusione continuava a tormentare Tallinn. E se un'intelligenza artificiale fosse stata in grado di simulare la mente umana? Questo non avrebbe significato, forse, che stavamo costruendo qualcosa di potenzialmente pericoloso? Il fondatore di Skype voleva approfondire l'argomento con Yudkowsky, ragion per cui aveva stilato una lista di domande, cercando di smontare alcune delle argomentazioni più catastrofistiche. Il modo migliore per capire se c'era del vero in quelle teorie era incontrare di persona il fondatore di LessWrong.

Così, avendo in programma di raggiungere San Francisco per una riunione, Tallinn inviò una e-mail a Yudkowsky chiedendogli se volesse incontrarlo per scambiare due chiacchiere. L'americano accettò e quando si sedettero in un caffè di Millbrae, non lontano dall'aeroporto internazionale di San Francisco, Tallinn iniziò a snocciolare le sue domande. Se l'AI era potenzialmente pericolosa, perché non costruirla semplicemente su macchine virtuali, isolandola da altri sistemi informatici? Di certo, questo avrebbe impedito all'AI di infiltrarsi nelle infrastrutture fisiche, magari per spegnere una rete elettrica o manipolare i mercati finanziari.

La risposta di Yudkowsky fu immediata. «Non sarebbe realmente virtuale», disse, sorseggiando il suo caffè. Gli elettroni potevano fluire in ogni direzione possibile, pertanto un sistema di AI abbastanza potente avrebbe comunque trovato il modo di interagire con l'hardware e modificarne la configurazione.

Tutto questo confermò i timori di Tallinn. Un giorno, pensò, l'AI avrebbe potuto sviluppare la propria infrastruttura e il proprio substrato informatico. Le implicazioni erano terribili, quanto a portata.

«Potrebbe terraformare e geo-ingegnerizzare il pianeta e forse persino il sole», dice oggi. Quando gli scienziati sostenevano che l'AI fosse solo matematica e che non c'era motivo di temerla, a Tallinn piaceva ribattere con l'analogia della tigre. «Potremmo sostenere che una tigre non è che un insieme di reazioni biochimiche e non c'è motivo di averne paura.» Ma una tigre è anche un insieme di atomi e cellule che, se non tenuto sotto controllo, può infliggere enormi danni. Allo stesso modo, l'AI potrebbe essere solo un'aggregazione di matematica avanzata e codice informatico; assemblata nel modo sbagliato, tuttavia, potrebbe essere incredibilmente pericolosa.

Quando, due anni più tardi, Tallinn si ritrovò ad ascoltare Hassabis alla conferenza di Oxford, era ormai un seguace convinto delle teorie sul rischio esistenziale dell'AI. Da quell'incontro al caffè, aveva divorato i saggi di Yudkowsky e si era immerso in un nuovo campo di ricerca, chiamato «allineamento dell'AI», nel quale scienziati e filosofi cercavano appunto di capire come «allineare» i sistemi di intelligenza artificiale agli obiettivi umani.

«Avevo mandato giù la pillola dell'allineamento», ricorda Tallinn. Al punto da credere ciecamente in alcuni degli scenari più estremi delineati da Yudkowsky.

Dopo qualche chiacchiera di rito, Tallinn tentò di capire se Hassabis fosse disposto a collaborare in maniera diretta. «Le andrebbe di fare una riunione su Skype, qualche volta?» gli chiese.

Hassabis e il facoltoso estone ebbero modo di riparlare, e alla fine Tallinn divenne uno dei primi investitori di DeepMind insieme a Peter Thiel. Il suo obiettivo non era solo fare soldi, ma anche tenere d'occhio i progressi di Hassabis e assicurarsi che non creasse involontariamente un'AI pericolosa e fuori controllo. Tallinn si riteneva un evangelista delle idee di Yudkowsky. Voleva sfruttare la sua credibilità di investitore affidabile per diffondere gli avvertimenti del filosofo ai costruttori di AI più promettenti al mondo.

«Eliezer è un autodidatta e non aveva molta influenza al di fuori della sua comunità ristretta», spiega Tallinn. «Ho pensato di poter iniziare a vendere quei ragionamenti a persone che non avrebbero ascoltato Eliezer, ma che avrebbero ascoltato me.»

Una volta diventato investitore, Tallinn spinse DeepMind a concentrarsi sulla sicurezza. Sapendo che Hassabis non condivideva i suoi timori per i rischi apocalittici dell'AI, fece pressione sull'azienda affinché assumesse un team che studiasse ogni maniera possibile per progettare l'AI in modo tale da mantenerla allineata ai valori umani e impedirle di uscire dai binari.

DeepMind stava per prendere a bordo un altro investitore con risorse ancora più ingenti e il medesimo intento di darle una direzione sicura. Nella Silicon Valley si sussurrava del coinvolgimento di Thiel in una promettente ma segretissima start-up con sede a Londra che cercava di sviluppare un'intelligenza artificiale generale. Tra i miliardari tecnologici della regione cui giunse la voce c'era anche Elon Musk. Nel 2012, due anni dopo aver cofondato DeepMind, Hassabis stava partecipando a una conferenza esclusiva in California organizzata da Thiel quando s'imbatté proprio in Musk.

«Ci siamo trovati subito in sintonia», racconta Hassabis. L'imprenditore britannico sapeva che quella poteva essere un'opportunità per raccogliere ulteriori fondi ed espandere la ricerca di DeepMind; inoltre, desiderava realmente poter visitare la fabbrica di razzi di Musk, che si stava affermando come magnate anticonformista con il sogno di portare gli esseri umani su Marte attraverso la sua azienda, SpaceX. Hassabis organizzò un incontro con Musk presso la sede di quest'ultima, a Los Angeles.

Di lì a poco, i due si trovarono seduti l'uno di fronte all'altro nella mensa aziendale, circondati da componenti di razzi, e finirono per discutere su quale dei loro progetti si sarebbe rivelato più importante per la storia dell'umanità: la colonizzazione interplanetaria o lo sviluppo di un'intelligenza artificiale superavanzata.

«Gli esseri umani dovranno essere in grado di rifugiarsi su Marte se mai l'AI dovesse andare fuori controllo», disse Musk, stando a un resoconto dell'incontro stilato da «Vanity Fair».

«Penso che l'AI sarebbe in grado di seguirci fin lì», rispose Hassabis, divertito. Musk, dal canto suo, non lo era affatto. Se Tallinn era stato influenzato dagli scritti

online di Yudkowsky, Musk era rimasto colpito dalle idee di un filosofo di Oxford di nome Nick Bostrom.

Bostrom aveva scritto un libro, *Superintelligenza*, che stava suscitando un certo scalpore tra gli esperti di AI e tecnologie di frontiera. Nel libro, Bostrom avvertiva che la costruzione di un'AI «generale» o estremamente potente avrebbe potuto portare a un esito disastroso per l'umanità, e non perché dovesse essere necessariamente malvagia o assetata di potere. Avrebbe potuto rivelarsi distruttiva anche solo nel cercare semplicemente di svolgere il proprio lavoro. Se per esempio le fosse stato assegnato il compito di produrre il maggior numero possibile di graffette, avrebbe potuto decidere di convertire tutte le risorse della Terra – esseri umani compresi – in graffette, essendo questo il modo più efficace per raggiungere l'obiettivo. Nei circoli legati all'AI, l'aneddoto aveva dato vita a un detto: bisognava evitare di lasciarsi «trasformare in graffette».

In ogni caso, anche Musk decise di investire in DeepMind. E benché questo consentisse finalmente a Hassabis di poter contare su una certa sicurezza finanziaria, non si trattava certo di una somma ingente. Stava comunque perseguendo qualcosa di così sperimentale e così folle che persino alcuni degli uomini più ricchi del pianeta esitavano a scommettere troppo sul successo dell'impresa. I loro finanziamenti, inoltre, portavano alcuni vincoli ideologici: Tallinn e Musk tenevano d'occhio il lavoro di DeepMind con un sospetto e una cautela insoliti per degli investitori. Volevano che DeepMind raggiungesse il successo finanziario, certo, ma evitando che crescesse troppo rapidamente o in modo tale da costituire un pericolo per l'umanità. E questo poneva Hassabis in una posizione scomoda: era grato loro per i soldi, ovviamente, ma non condivideva gli scenari catastrofici che dipingevano.

Quella sensazione di sicurezza finanziaria, tuttavia, non durò a lungo. Hassabis e Suleyman faticavano a guadagnare abbastanza per coprire i costi necessari a pagare le migliori menti nel settore dell'AI, e alcune delle loro idee per generare entrate erano alquanto incoerenti. Provarono a lanciare un sito web che utilizzasse il *deep learning* – il genere di apprendimento automatico in cui DeepMind si era specializzata in un primo momento – per impartire consigli di moda e raccomandare capi d'abbigliamento. Poi Hassabis chiese ad alcuni membri del team, con cui aveva già collaborato in Elixir e che ora lavoravano per DeepMind, di progettare un videogioco. Secondo un ex dipendente, gli ingegneri idearono un'avventura spaziale in cui un equipaggio di astronauti doveva raggiungere la Luna a bordo di un razzo. Stavano per lanciare il gioco come un'app per iPhone quando a Hassabis si presentò una nuova opportunità, qualcosa che gli avrebbe dato il supporto finanziario necessario per rendere l'AGI una realtà: un'offerta da parte di Facebook.

Mark Zuckerberg era in piena fase di acquisizioni. Circa un anno prima, aveva rilevato Instagram per un miliardo di dollari, con quella che si sarebbe dimostrata una mossa magistrale di consolidamento nei social media. E pochi mesi dopo avrebbe sborsato l'astronomica cifra di 19 miliardi di dollari ai fondatori di WhatsApp. Era

pronto a spendere qualsiasi somma per espandere l'impero di Facebook, e l'intelligenza artificiale avrebbe giocato un ruolo importante nell'impresa. Facebook guadagnava circa il 98 per cento dei suoi profitti vendendo pubblicità ma, per aumentare il numero di inserzioni e continuare a crescere, a Zuckerberg serviva che le persone trascorressero sempre più tempo sulle sue piattaforme. E gli scienziati AI di DeepMind avrebbero potuto fornire un contributo prezioso in tal senso. Con sistemi di raccomandazione più sofisticati, in grado di analizzare a fondo i dati personali degli utenti, gli algoritmi alla base di Facebook e Instagram avrebbero potuto mostrare immagini, post e video sempre più mirati, in modo da mantenere le persone sempre più incollate allo schermo.

Zuckerberg offrì a Hassabis 800 milioni di dollari per DeepMind, senza contare il bonus che i fondatori di start-up solitamente ricevevano rimanendo presso l'azienda acquisita per quattro o cinque anni. Era una cifra generosa: più denaro di quanto Hassabis avesse mai sognato. E questo lo pose davanti a un bivio. Fino a quel momento, i finanziamenti erano arrivati da persone che gli chiedevano di usare la massima cautela nello sviluppo dell'AI. Ora, invece, il denaro poteva giungere da qualcuno che intendeva accelerare il processo. Dopotutto, il motto di Facebook era «Muoviti in fretta e rompi le cose».

Hassabis e Suleyman ne discussero a lungo. L'AGI sarebbe stata più potente di quanto lo stesso Zuckerberg potesse immaginare, ed entrambi ritenevano necessario prendere le giuste misure per impedire che un grande acquirente orientasse l'AI verso una direzione potenzialmente dannosa. Non potevano semplicemente limitarsi a far firmare a Facebook un contratto in cui prometteva di non abusare dell'AGI. Ripensando alle esperienze pregresse con organizzazioni non profit, Suleyman disse a Hassabis e Legg che avevano bisogno di una sorta di struttura di governance che monitorasse da vicino Facebook e garantisse un uso responsabile della tecnologia di DeepMind.

Di solito, le aziende quotate in Borsa hanno un consiglio di amministrazione il cui compito è rappresentare gli interessi degli azionisti. Il consiglio si riunisce ogni trimestre per esaminare le azioni dell'azienda e assicurarsi che tutto venga fatto nel modo giusto per accrescerne il valore, anziché farlo scendere. Suleyman disse ai cofondatori che DeepMind avrebbe dovuto avere un tipo di consiglio differente, adatto a una tecnologia trasformativa come l'AI. Invece di concentrarsi sul denaro, il suo compito sarebbe stato garantire che DeepMind sviluppasse l'AI nel modo più sicuro ed etico possibile. Pur non essendo inizialmente convinti, alla fine Hassabis e Legg accettarono la proposta.

Hassabis tornò da Zuckerberg e gli disse che, se avessero venduto, DeepMind avrebbe dovuto istituire un comitato per l'etica e la sicurezza, con un'autorità legale indipendente per controllare qualsiasivoglia AI superintelligente sviluppata in futuro da DeepMind. Zuckerberg si oppose alla richiesta. Il suo intento era far crescere il business pubblicitario di Facebook e «connettere il mondo» attraverso le sue

piattaforme social, non quello di gestire una società di AI separata con una serie di protocolli etici e una propria missione. Le trattative saltarono.

Hassabis disse ai suoi dipendenti che DeepMind avrebbe conservato la propria autonomia per altri vent'anni. In privato, però, era stanco di dover raccogliere continuamente fondi e frustrato dal fatto di poter dedicare solo una frazione del suo tempo alla ricerca effettiva. Dopo aver rifiutato un'offerta spropositata da parte di Zuckerberg, era difficile ignorare quanto avrebbe potuto guadagnare vendendo a una compagnia della Silicon Valley, specialmente ora che Big Tech sembrava improvvisamente sbavare sull'AI. I dirigenti delle più grandi aziende della Silicon Valley, tra cui uno o due miliardari, chiamavano regolarmente i ricercatori di DeepMind per reclutarli, anche perché molti dei dipendenti dell'azienda erano esperti in *deep learning*, una branca a lungo considerata marginale nel campo dell'intelligenza artificiale, almeno fino a poco tempo prima.

Il punto di svolta era arrivato nel 2012. Una docente di AI di Stanford, Fei-Fei Li, aveva istituito una competizione annuale per accademici, chiamata ImageNet, nel corso della quale i ricercatori presentavano modelli di AI che cercavano di riconoscere visivamente immagini di gatti, mobili, automobili e altro ancora. Quell'anno, il team del ricercatore Geoffrey Hinton aveva utilizzato l'apprendimento profondo per creare un modello molto più preciso dei precedenti, con risultati stupefacenti. A un tratto, ecco che tutti volevano assumere esperti in questa teoria dell'AI basata sul *deep learning* e ispirata al modo in cui il cervello riconosce i pattern.

Si trattava di un campo ristretto, con appena una manciata di esperti, dice Legg. «Parecchi dei quali lavoravano per noi.» Hassabis li pagava circa 100.000 dollari l'anno, ma giganti tecnologici come Google e Facebook avrebbero sborsato molto di più. «Ci chiamavano persone davvero famose per offrire tre volte il loro stipendio», ricorda Legg. Zuckerberg era proprio una di queste, secondo un ex dipendente di DeepMind. «Dovevamo vendere, altrimenti ci avrebbero fatto a pezzi.» E Hassabis, ansioso com'era di costruire l'AGI prima degli altri, non poteva aspettare che aziende tecnologiche con più risorse lo precedessero.

All'improvviso arrivò un'altra offerta di acquisto, questa volta da parte di un suo investitore, Elon Musk. Secondo una persona informata, il miliardario si proponeva di pagare l'acquisto con azioni di Tesla, l'azienda di auto elettriche che guidava ormai da cinque anni. Nei panni di investitore, Musk si era rivelato poco invadente, e solo di rado aveva fatto visita a Hassabis. Nonostante le crescenti preoccupazioni circa i pericoli legati all'AI, il miliardario aveva anche ben presenti i propri obiettivi commerciali. Voleva che le Tesla fossero le prime auto al mondo a utilizzare con successo la tecnologia di guida autonoma, e a tale scopo aveva bisogno di esperti all'avanguardia nel campo dell'AI. Ora poteva assicurarsi un'intera élite di esperti acquistando DeepMind.

Ancora una volta, però, i fondatori dell'azienda erano scettici. Essere pagati in azioni Tesla non sembrava così allettante. Inoltre, l'idea che uno come Musk

prendesse il controllo dell'AGI li metteva a disagio. Benché stesse appena iniziando a guadagnare fama mainstream come imprenditore visionario, nei circoli tecnologici Musk aveva la nomea di manager bizzoso, tanto da licenziare dipendenti senza preavviso e cacciare persino il cofondatore di Tesla.

Per quanto i fondatori di DeepMind apprezzassero il suo investimento e la sua rete di contatti, diffidavano del suo comportamento imprevedibile. Così, rifiutarono anche la sua offerta, senza rendersi conto di quanto Musk detestasse sentirsi dire di no e di quanto quella decisione avrebbe potuto ritorcersi contro. Di lì a poco, però, Hassabis ricevette un'altra e-mail. Questa volta da Google.

5. PER L'UTOPIA, PER I SOLDI

Il messaggio proveniva da un dirigente della sede di Google, a più di ottomila chilometri di distanza, nella soleggiata Mountain View, in California. Aprendolo sul suo computer, a Londra, Hassabis trovò un invito a incontrare Larry Page, il CEO di Google. Page aveva cofondato Google con un altro dottorando di Stanford, Sergey Brin, nel 1998. I due volevano facilitare le ricerche su internet, e lo avevano fatto creando un algoritmo denominato PageRank, che classificava le pagine web in base a rilevanza e interconnessione. Partendo dal garage di un amico a Menlo Park, in California, avevano finito per strutturare una delle più grandi aziende tecnologiche al mondo.

Quanto al modo in cui Google guadagnava, non era particolarmente high-tech o innovativo: era diventata una gigantesca azienda pubblicitaria, come Facebook. La maggior parte dei profitti di Google proveniva dal monitoraggio delle informazioni personali al fine di raggiungere gli utenti con pubblicità mirate tramite la ricerca, YouTube, Gmail e milioni di siti web e app che usavano la Rete Display di Google.
ESCLUSIVA EUREKADDL

C'era qualcosa di leggermente inquietante in tutto questo, per uno come Hassabis che voleva usare l'AI per aiutare il mondo. D'altro canto era consapevole che, se non avesse accettato l'offerta, Google avrebbe potuto soffiargli il personale e magari costruire l'AGI senza di lui. Google aveva già centinaia di ingegneri che lavoravano sull'AI, e proprio per questo Hassabis si persuase ad accettare l'invito.

Quando incontrò Page, ebbe la sensazione di parlare con uno spirito affine. Davanti a lui c'era un introverso laureato in matematica dalle folte sopracciglia scure che indossava camicie casual e bermuda. Per un'intera carriera dedicata alla costruzione di Google, Page aveva coltivato il sogno di creare un'AI performante. «Mi disse di aver sempre pensato a Google come a un'azienda di AI, sin da quei primi giorni del 1998, nel garage», ricorda Hassabis.

In parte, la questione era personale: suo padre era stato docente di intelligenza artificiale e scienze informatiche fino alla morte, avvenuta nel 1996. Questo faceva di lui una sorta di tecnologo dell'AI di seconda generazione. Page ammirava la serietà con cui Hassabis stava costruendo l'AGI, e non considerava affatto strampalate le sue idee. Aveva già dato il via a un progetto interno a Google per sviluppare un'AI che imitasse quella umana, un'iniziativa che in futuro avrebbe finito per alimentare una forte rivalità con lo stesso Hassabis.

Il progetto di Page, di cui all'epoca Hassabis non era a conoscenza, si chiamava Google Brain. Era nato da una proposta di Andrew Ng, un professore di Stanford dall'eloquio pacato che voleva costruire sistemi di AI più avanzati all'interno di Google. Nel 2011, qualche anno prima che Google contattasse DeepMind, Ng aveva inviato a Page un documento di quattro pagine dal titolo *Neuroscience-Informed Deep Learning* («Apprendimento profondo basato sulle neuroscienze»). Con questa mossa, sperava che il CEO di Google approvasse un progetto mirato a realizzare sistemi di AI «per uso generico», lo stesso obiettivo su cui Hassabis lavorava in Inghilterra.

Venne fuori che Ng e Hassabis approcciavano la questione con un metodo simile, cercando entrambi ispirazione nelle neuroscienze. Nella sua proposta, il professore di Stanford aveva spiegato a Page che il suo intento era di ricreare «approssimazioni sempre più accurate di piccole parti del cervello dei mammiferi».

Anche per uno come Ng, che era già una figura di spicco nel settore e lavorava per una delle università più prestigiose al mondo, l'idea di costruire un'AGI era controversa. «I miei amici mi consigliavano di lasciar perdere. Dicevano: "Non gioverà alla tua carriera"», ricorda.

In un certo senso, avevano ragione. Sotto il profilo scientifico, l'ossessione di Ng e Hassabis per il cervello umano presentava alcuni problemi. In teoria, usare la nostra materia grigia come modello per l'AI era una scelta logica, ma copiare ciò che troviamo in natura non sempre funziona. Basti pensare ai primi tentativi di realizzare macchine volanti e agli inventori che costruivano congegni che imitavano la meccanica del volo degli uccelli e che finivano puntualmente per schiantarsi al suolo. Altri scienziati informatici si erano già ritrovati in un vicolo cieco tentando di copiare troppo fedelmente il cervello. Nel 2013, il neuroscienziato Henry Markham affermò in un TED Talk di aver scoperto il modo di simulare un intero cervello umano su supercomputer e ci sarebbe riuscito nel giro di un decennio. Dieci anni più tardi, il suo Human Brain Project, costato già oltre un miliardo di dollari, poteva dirsi in gran parte fallito.

Negli anni successivi, Ng, Hassabis e altri scienziati nel campo dell'AI avrebbero preso coscienza di quanto fosse difficile emulare il cervello, avendone ancora una conoscenza così incompleta, dalle funzioni dei neuroni alla dinamica delle regioni cerebrali. Sapevamo già che nel nostro cranio ci sono circa novanta miliardi di neuroni che si attivano continuamente, ma non avevamo alcuna idea di come venissero elaborate le informazioni.

«Con il senno di poi, provare a rimanere così fedeli alla biologia è stato un errore», afferma Ng. Tuttavia, la ricerca del professore si rivelò efficace nell'ampliare le sue reti neurali.

Una rete neurale è un tipo di software che viene costruito tramite un processo di addestramento ripetuto grazie a enormi quantità di dati. Una volta addestrata, una rete neurale può riconoscere volti, prevedere mosse di scacchi o consigliarvi il prossimo

film su Netflix. Chiamata anche «modello», una rete neurale è spesso composta da molti strati e nodi che elaborano informazioni in modo vagamente simile ai neuroni del nostro cervello. Più il modello viene addestrato, più quei nodi migliorano la propria capacità di prevedere o riconoscere le cose.

Ng aveva scoperto che questi modelli potevano svolgere più compiti se disponevano di un maggior numero di nodi, strati e dati su cui allenarsi. Anni dopo, anche OpenAI avrebbe fatto una scoperta simile sull'importanza di potenziare questi elementi chiave. Nei suoi esperimenti a Stanford, Ng aveva notato che i suoi modelli di *deep learning* funzionavano molto meglio quando erano costruiti in scala più grande. Entusiasmato dai risultati, aveva inviato la sua proposta di quattro pagine a Page, suggerendo di poter costruire «simulazioni cerebrali su larga scala» come primo passo verso un'«AI di livello umano».

Innamoratosi dell'idea, Page prese a bordo Ng perché guidasse il progetto di ricerca più all'avanguardia di Google fino a quel momento. Pochi anni dopo, però, Google Brain non sembrava più così orientato verso la costruzione di un'AGI. Piuttosto, stava aiutando Google a ottimizzare il suo business pubblicitario mirato – rendendo gli annunci ancora più spaventosamente accurati per gli utenti, migliorando la capacità di prevedere su cosa avrebbero cliccato questi ultimi – e a incrementarne i ricavi. Ng ammette che non era ciò per cui aveva inviato la sua proposta a Page. «Non è la cosa più entusiasmante a cui abbia lavorato», confessa.

Ciò a cui ambiva veramente Ng con la sua ricerca era affrancare l'umanità dalla fatica mentale, nello stesso modo in cui la rivoluzione Industriale ci aveva liberati dal lavoro fisico incessante. Credeva che sistemi di AI più avanzati avrebbero fatto lo stesso per i lavoratori professionisti, «permettendo a noi tutti di dedicarci a compiti più stimolanti e di alto livello sotto il profilo intellettuale».

Ma era nelle modalità con cui perseguiva questo obiettivo che Ng differiva da Hassabis. Mentre l'imprenditore britannico voleva quanta più indipendenza possibile dal colosso pubblicitario, il professor Ng era ben felice di lavorare nel ventre della bestia. In tal senso, aveva reso un enorme favore a Hassabis. Stabilendosi all'interno della casa madre di Google, la sua ricerca era già destinata a contribuire al business pubblicitario dell'azienda, il che evitava a DeepMind di doverlo fare nell'immediato.

Quando Google contattò DeepMind, alla fine del 2013, i ricercatori di Ng erano già stati assorbiti dallo sviluppo di sofisticati modelli di AI con cui potenziare gli strumenti pubblicitari di Google, distogliendoli dall'obiettivo più ambizioso di Ng, ovvero quello di creare un'AI onnipotente capace di liberare l'umanità dai lavori poco stimolanti. Mentre volava a Londra per negoziare l'acquisto di DeepMind, Page sapeva di poter investire una parte del denaro di Google per qualcosa di un po' più visionario.

I fondatori di DeepMind accolsero il miliardario di Google nel loro ufficio di Londra, con una presentazione sui progressi dell'azienda fino a quel momento, secondo quanto Cade Metz, corrispondente del «New York Times», riporta nel suo

Costruire l'intelligenza. Hassabis spiegò come il suo team avesse sviluppato una nuova tecnica chiamata *reinforcement learning* («apprendimento per rinforzo») per addestrare un sistema di AI a padroneggiare il gioco retrò Atari *Breakout*. Nel gioco, bisogna spedire una pallina contro una parete di mattoni utilizzando una paletta che si muove da un lato all'altro. In circa due ore, il sistema aveva imparato a colpire la pallina nel punto giusto, in modo che aprisse un varco nello spazio stretto oltre la fila superiore di mattoni, distruggendone un gran numero in un colpo solo. Page ne rimase impressionato.

L'apprendimento per rinforzo non era molto diverso dal modo in cui si premia un cane con dei bocconcini ogni volta che si siede a comando. Nell'addestramento dell'AI, si premia analogamente il modello, magari con un segnale numerico come un +1, per dimostrare che un certo risultato è positivo. Attraverso ripetuti tentativi ed errori, e giocando centinaia di partite, il sistema imparava cosa funzionava e cosa no. Era un'idea di una elegante semplicità confezionata in un codice informatico estremamente sofisticato.

Poi Legg illustrò a Page la direzione verso cui puntava questa ricerca: l'applicazione di queste tecniche al mondo reale. Così come il loro sistema aveva imparato a padroneggiare un videogioco, avrebbero potuto insegnare a un robot a orientarsi in una casa o a un agente autonomo a destreggiarsi nella lingua inglese. Erano queste le applicazioni in cui le scoperte di DeepMind e l'AGI stessa avrebbero avuto il maggior impatto. E ciò convinse Page e il suo team.

Il CEO di Google guidò le trattative con Hassabis e i suoi cofondatori sapendo che avevano già rifiutato un'offerta sostanziosa da parte di Facebook. Stava per scoprirne il motivo. Hassabis espose le due condizioni fondamentali per la vendita. Per prima cosa, lui e i cofondatori non volevano che Google usasse la tecnologia di DeepMind per scopi militari, che si trattasse di dirigere droni autonomi o armi, o di supportare soldati sul campo. Per loro, questi erano confini etici invalicabili.

In secondo luogo, volevano che i dirigenti di Google firmassero un accordo di etica e sicurezza. Redatto da avvocati londinesi, il contratto cedeva il controllo di qualsiasi tecnologia di intelligenza artificiale generale sviluppata in futuro da DeepMind a un comitato etico istituito da Hassabis e Suleyman. Entrambi avevano ancora solo una vaga idea di chi dovesse farne parte, ma volevano che avesse il controllo legale assoluto sull'AI avanzata che avrebbero costruito.

«Se fossimo riusciti nell'impresa, sarebbe stato necessario gestire [l'AGI] con attenzione», dice Hassabis riguardo al comitato. «Data la sua versatilità, poteva trattarsi di una delle tecnologie più potenti di tutti i tempi, perciò all'epoca volevamo la certezza di trattare con persone che se ne accollassero seriamente le responsabilità.»

Non sorprende che ci siano voluti mesi di complesse trattative per convincere Google ad accettare la medesima condizione che aveva frenato Facebook. L'acquisto di DeepMind consegnava a Page la prima azienda in grado di costruire un'AGI. E lui

sapeva che se questo comitato etico avesse avuto il controllo legale sulla tecnologia in questione, sarebbe stato molto più complicato, per Google, trarne profitto; tuttavia, alla fine, la sua visione idealista prevalse. Avrebbero trovato un modo per far funzionare le cose. Di conseguenza, accettò le condizioni di DeepMind circa il comitato etico.

L'AGI non necessitava di una gestione attenta solo in relazione ai possibili scenari futuri legati ai giganti aziendali. Era anche al centro di diverse ideologie che avrebbero potuto indirizzare la tecnologia in diverse direzioni. Hassabis ne aveva avuto un assaggio da investitori come Peter Thiel, che voleva accelerarne lo sviluppo, e da Jaan Tallinn, che invece temeva l'innesco di una sorta di apocalisse.

Il potenziale sconvolgente dell'AI esercitava una fascinazione quasi mistica su individui animati da convinzioni forti su come dovesse essere utilizzata. Negli anni a venire, queste forze ideologiche si sarebbero scontrate con gli innovatori e i monopoli aziendali che si contendevano il controllo dell'AGI, diventando un pericolo imprevedibile per la tecnologia. Avrebbero spinto, per esempio, Sam Altman fuori da OpenAI e, paradossalmente, rafforzato gli sforzi commerciali delle aziende, dipingendo un quadro apocalittico del potere dell'AI che finiva per rendere il software più attraente per le imprese. A contatto con il mondo degli affari e del profitto, sempre più costruttori di AI si ritrovavano a seguire devotamente dottrine diverse, da quella permeata dall'ansia di realizzare quanto prima l'AI per dare vita a un'utopia a quella che fomentava la paura che potesse provocare l'Armageddon.

In quanto pensatore strategico che procedeva sempre con i piedi di piombo, Hassabis si trovò in gran parte fuori da questi principi ideologici in conflitto, grazie anche all'obiettivo che aveva in testa, ovvero giungere a scoperte straordinarie e forse persino divine tramite l'AGI. Suleyman era anche più preoccupato per i problemi sociali che l'AI poteva causare prima di quel momento. Dei tre cofondatori, Shane Legg era quello maggiormente in linea con le ideologie più radicali legate alla ricerca dell'AGI, inclusa una che, stando alle parole dei suoi ex colleghi, covava ormai da decenni. Nota come transumanismo, l'idea aveva radici controverse e una storia che contribuiva a spiegare perché i costruttori di AI talvolta trascurassero gli effetti collaterali sgradevoli, e più attuali, della tecnologia.

La premessa di base del transumanismo è che noi umani siamo al momento una specie di qualità inferiore. Con le appropriate scoperte scientifiche e tecnologiche, un giorno potremmo evolverci oltre gli attuali limiti fisici e mentali in una specie più dotata. Saremo più intelligenti e creativi, e vivremo più a lungo. Potremmo anche riuscire a fondere le nostre menti con i computer e a esplorare la galassia.

L'idea centrale risale al periodo tra gli anni Quaranta e Sessanta del secolo scorso, quando un biologo evoluzionista di nome Julian Huxley aderì alla British Eugenics Society, diventandone il presidente. Il movimento eugenetico sosteneva che gli esseri umani dovessero migliorarsi attraverso la selezione della riproduzione, e si diffuse nelle università britanniche e tra le classi intellettuali e altolocate del paese. Lo stesso

Huxley proveniva da una famiglia aristocratica (suo fratello Aldous scrisse *Il mondo nuovo*) e riteneva che l'élite della società fosse geneticamente superiore. Le persone delle classi inferiori andavano eliminate come una messe cattiva e sottoposte a sterilizzazione forzata. «Si stanno riproducendo troppo in fretta», scriveva Huxley.

Quando i nazisti fecero propria l'ideologia eugenetica, Huxley decise che il movimento aveva bisogno di un nuovo nome. Coniò così il termine *transumanismo* in un saggio in cui affermava che, oltre alla riproduzione selettiva, l'umanità avrebbe anche potuto «trascendere sé stessa» attraverso la scienza e la tecnologia. Il movimento acquistò slancio negli anni Ottanta e Novanta, quando il nascente campo dell'intelligenza artificiale suggerì una possibilità allettante: forse gli scienziati avrebbero potuto potenziare la mente umana attraverso una fusione con macchine intelligenti.

L'idea si cristallizzò nel concetto di singolarità, quel momento collocato nel futuro in cui l'AI e la tecnologia avrebbero raggiunto un livello talmente avanzato da provocare un cambiamento drammatico e irreversibile nel genere umano, fondendolo con le macchine e potenziandolo. Si trattava di una prospettiva che sin dall'adolescenza aveva affascinato Legg, così come il facoltoso finanziatore di DeepMind Peter Thiel. Tra i tecnologi, l'attrazione per questa utopia era tale che alcuni di loro, come Altman e Thiel, si erano registrati presso diverse aziende specializzate nella criogenizzazione del cervello o dell'intero corpo, nel caso in cui non fossero riusciti a fondersi con le macchine prima della morte. «Non mi aspetto necessariamente che funzioni», disse Thiel alla giornalista Bari Weiss nel suo podcast. «Ma credo che sia il genere di cosa che dovremmo almeno provare a fare.»

Il problema con alcune di queste idee era che, nel corso degli anni, i loro seguaci divennero sempre più fanatici. I cosiddetti «accelerazionisti dell'AI», per esempio, credevano che gli scienziati avessero l'imperativo morale di lavorare il più velocemente possibile per sviluppare un'AGI capace di creare un paradiso postumano, una sorta di estasi tecnologica per nerd. Se fosse stata realizzata durante l'arco della loro esistenza, avrebbero potuto vivere per sempre. Tuttavia, accelerare lo sviluppo dell'AI poteva anche significare prendere scorciatoie e generare una tecnologia potenzialmente dannosa per determinati gruppi di persone o capace di sfuggire al controllo.

Altri, invece, si posizionavano sulla sponda opposta, vedendo nell'AI una sorta di minaccia demoniaca da fermare a tutti i costi. Eliezer Yudkowsky, il barbuto libertario che aveva contribuito a radicalizzare Jaan Tallinn davanti a un caffè, era uno dei leader di questo movimento ideologico, al quale diede sempre più slancio attraverso il già citato sito LessWrong. Quando Google acquistò DeepMind, nel 2014, centinaia di persone, tra cui numerosi ricercatori di AI, stavano partecipando a dibattiti di carattere filosofico sul sito, discutendo di come evitare che una superintelligenza, in futuro, finisse per annientare il genere umano. LessWrong era diventato l'hub online delle paure apocalittiche legate all'AI, e alcuni articoli sottolineavano come

ormai avesse tutte le caratteristiche di un moderno culto millenarista. Quando un partecipante suggerì l'ennesimo possibile scenario in cui l'AI avrebbe potuto distruggere l'umanità, lo stesso Yudkowsky reagì con veemenza, attaccandolo pubblicamente con un messaggio tutto in maiuscolo e infine espellendolo dal gruppo.

Con il tempo, i cosiddetti *doomers* dell'AI guadagnarono abbastanza sostegno tra i ricchi imprenditori tecnologici da ottenere finanziamenti con cui avviare aziende e plasmare le politiche governative. Il sito di Yudkowsky divenne così influente che molti dei suoi lettori più assidui finirono per entrare in OpenAI.

Ma forse le ideologie più inquietanti che iniziavano a diffondersi intorno all'AGI erano quelle che miravano a creare una specie umana quasi perfetta in forma digitale. L'idea era stata in parte resa popolare dal già citato *Superintelligenza* di Bostrom, libro che ebbe un impatto paradossale sul campo dell'AI: da un lato, alimentando i timori sulla possibile distruzione del genere umano a opera di un'AI superintelligente e fuori controllo che avrebbe potuto sterminarci senza nemmeno volerlo, trasformandoci in graffette; dall'altro, descrivendo la possibilità di un'utopia gloriosa verso cui avrebbe potuto introdurci l'AI, se fosse stata costruita nel modo giusto.

Uno degli aspetti più affascinanti di questa visione, secondo Bostrom, era la creazione di «postumani» dotati di capacità enormemente superiori a quelle degli umani «normali» e destinati a vivere in substrati digitali. Secondo questa utopia digitale, gli esseri umani avrebbero potuto esplorare ambienti che sfidavano le leggi della fisica, decidere autonomamente la propria fine o esplorare mondi fantastici. Avrebbero potuto rivivere ricordi preziosi, creare nuove avventure o sperimentare diverse forme di coscienza. Le interazioni sarebbero diventate più profonde, poiché questi nuovi esseri umani avrebbero potuto condividere pensieri ed emozioni in maniera diretta, portando a connessioni più intime e significative.

Queste idee erano irresistibili per alcuni esponenti della Silicon Valley, convinti com'erano che una simile realtà fosse raggiungibile grazie ad algoritmi appropriati. Presentando un futuro che poteva somigliare tanto al paradiso quanto all'inferno, Bostrom alimentò una convinzione che avrebbe spinto i costruttori di AI della Silicon Valley, come Sam Altman, a gareggiare per sviluppare l'AGI prima di Demis Hassabis a Londra. Dovevano arrivare per primi, perché *soltanto loro* potevano farlo in sicurezza. Se qualcun altro fosse riuscito a precederli, senza un adeguato allineamento ai valori umani, c'era il rischio di annientare non solo i miliardi di umani viventi, ma anche le potenziali migliaia di miliardi di umani-digitali del futuro. In sostanza, avremmo perso la possibilità di vivere in un nirvana tecnologico. Nel frattempo, le teorie di Bostrom avrebbero avuto anche effetti collaterali pericolosi, distogliendo l'attenzione dallo studio dei modi in cui l'intelligenza artificiale poteva danneggiare le persone già nel presente.

Mentre queste ideologie prendevano piede e s'intrecciavano con le trattative tra DeepMind e Google, emergeva una scomoda realtà: per le aziende tecnologiche, trovare una forma di gestione responsabile dell'AI diventava sempre più difficile.

Obiettivi contrastanti erano destinati a collidere, spinti da un fervore quasi religioso da un lato e da una fame insaziabile di crescita economica dall'altro.

Per il momento, e in virtù dei motivi personali che lo spingevano a realizzare l'AGI, Hassabis rimaneva distante da queste battaglie ideologiche. Operando in Inghilterra, lontano dalla bolla della Silicon Valley, si era circondato di un team di scienziati e ingegneri straordinariamente brillanti, destinato tra l'altro a crescere. Era determinato a risolvere l'enigma dell'AGI nel giro di cinque anni, un'impresa che – secondo chi lavorava con lui – avrebbe potuto garantirgli persino un Nobel. Né gli importava il fatto che un colosso aziendale potesse inglobarlo: una volta costruita l'AGI, il concetto stesso di economia sarebbe diventato obsoleto, e né DeepMind né Google avrebbero più dovuto preoccuparsi di fare soldi. L'AI avrebbe risolto anche quel problema.

Quando l'accordo fu finalmente firmato e il comitato etico aggiunto al contratto di acquisizione, Google acquistò DeepMind per 650 milioni di dollari. Era una cifra inferiore a quella che i fondatori avrebbero ottenuto da Zuckerberg, ma comunque enorme per un'azienda tecnologica britannica, e includeva un elemento cruciale: la garanzia che il controllo dell'AGI non sarebbe finito nelle mani di una grossa corporation.

L'afflusso di denaro da parte di Google consentì a Hassabis di reclutare ancora più ricercatori di talento. Benché alcuni dipendenti non fossero entusiasti di «svendersi» a Google, molti altri erano euforici per i corposi aumenti di stipendio e le allettanti stock option che rendevano meno appetibile la concorrenza. Ora, invece di temere che Facebook o Amazon gli portassero via il personale, Hassabis poteva fare il contrario: reclutare *i loro* migliori talenti e attirare alcune delle menti più brillanti del mondo accademico con stipendi da capogiro. In ogni caso, guidando l'azienda verso lo sviluppo di tecnologie sempre più avanzate, Hassabis mantenne la cultura improntata alla segretezza di DeepMind. Il suo sito web, per esempio, rimase una semplice pagina bianca con al centro un logo circolare. Il laboratorio di AI era così avvolto nel mistero che chiunque si candidasse per un lavoro presso la sede londinese non riceveva nemmeno l'indirizzo via e-mail: un rappresentante gli andava incontro alla stazione ferroviaria di King's Cross e lo accompagnava a piedi fino all'ufficio.

«Nei colloqui di lavoro, i fondatori erano persuasivi, specie Suleyman», racconta un ex dirigente. «Era estremamente carismatico e trasmetteva l'idea che si trattasse dell'opportunità irripetibile di far parte di qualcosa destinato a cambiare il mondo.»

Accademici e funzionari pubblici con oltre dieci anni di carriera alle spalle, che avrebbero potuto facilmente accedere a ruoli ben retribuiti nel settore privato, uscivano dopo venti minuti di conversazione con Suleyman convinti di dover contribuire alla costruzione dell'AGI. «Spiegava che la rivoluzione si sarebbe basata su una matematica migliore», aggiunge l'ex dirigente. Hassabis e Suleyman sostenevano di assumere «i migliori matematici e fisici del mondo». E ora, in Google, avevano accesso ai supercomputer più avanzati e a un'enorme quantità di dati con cui

addestrare i loro modelli di AI.

Circa la metà dei nuovi assunti di DeepMind proveniva ormai dal mondo accademico, e non riusciva a credere alla propria fortuna. Schiacciati per anni tra gli schedari e costretti a mendicare fondi, si ritrovavano ora in uffici scintillanti, circondati da ristoranti cosmopoliti e giardini, con computer ultraveloci e risorse pressoché illimitate. La parte migliore? DeepMind faceva in modo che non sembrasse di lavorare per un colosso della pubblicità. Si conducevano ricerche in un'organizzazione scientifica prestigiosa, pubblicando articoli su riviste peer-reviewed come «Science» e «Nature» e affrontando alcune delle sfide più urgenti per il pianeta. Era il meglio di entrambi i mondi, se mai una cosa simile fosse stata possibile.

Sul lungo termine, non lo era. Ma stipendi a sei cifre e benefit incredibili facevano dimenticare ai dipendenti di DeepMind quanto fosse strano essere pagati così generosamente da Google solo per rendere il mondo un posto migliore. A volte, però, questa incongruenza affiorava, come quando ex colleghi del polveroso mondo accademico o della pubblica amministrazione chiedevano di visitare gli uffici.

«Mi sentivo in imbarazzo», confessa un ex dipendente, passato a DeepMind dopo aver lasciato un incarico universitario. Quando i suoi ex colleghi gli domandavano di vedere il suo nuovo ufficio, cercava di dissuaderli, proponendo invece di incontrarli in un ristorante nei dintorni. Anche quello sarebbe stato un ambiente più sobrio rispetto alla mensa di DeepMind, che offriva buffet degni di un cinque stelle a Dubai. «Sembrava un mondo a parte», aggiunge. «Era davvero assurdo.»

I ricercatori erano serviti e riveriti come rockstar. Un giorno, uno di loro scrisse al servizio di supporto interno – utilizzato in genere per rimborsi spese o richieste di visto – suggerendo che sarebbe stato più efficiente servire tutte le fragole già private delle foglioline. Due giorni dopo, al buffet comparvero ciotole di fragole pulite alla perfezione.

La visione dietro la costruzione dell'AGI veniva costantemente riproposta ai dipendenti, e Hassabis andava ripetendo che, al ritmo con cui progredivano, l'obiettivo finale distava ormai solo cinque anni. A detta di chi lavorava in DeepMind, Hassabis era un maestro nel dipingere una visione ispirata sul futuro dell'azienda. Durante i ritiri aziendali, lui e Suleyman tenevano presentazioni che somigliavano più a comizi motivazionali che a spiegazioni dettagliate di piani operativi. I fondatori, infatti, spesso evitavano di scendere nei dettagli sulle strategie adottate.

«Era tutto molto orientato alla visione, alla missione da compiere», racconta un ex dipendente. «Demis e Mustafa erano affabulatori straordinari. Si bilanciavano incredibilmente bene.» Hassabis era la mente, l'uomo che leggeva articoli scientifici fino a notte fonda, che discuteva per ore di metodologie con i suoi ricercatori di punta e tendeva a evitare il personale di rango inferiore che non aveva un dottorato. Era stato lui a plasmare la cultura profondamente gerarchica di DeepMind, basata in gran

parte sul prestigio accademico. Suleyman, invece, era il visionario carismatico, capace di delineare il futuro per cui tutti stavano lavorando. Un ex membro del team lo descrive come il pifferaio magico di DeepMind. Shane Legg, il più accademico del trio, si teneva più sullo sfondo. «Shane era il più taciturno», osserva l'ex dipendente.

Hassabis credeva così fermamente nel potere trasformativo dell'AGI da ripetere ai dipendenti di DeepMind che, nel giro di cinque anni, non avrebbero più dovuto preoccuparsi del denaro, perché l'AGI avrebbe reso l'economia obsoleta. Questa divenne presto la convinzione dei vertici. «Avevano bevuto il loro stesso Kool-Aid», commenta un ex dirigente. Erano convinti di essere sul punto di creare la tecnologia più importante che l'umanità avesse mai conosciuto.

Dietro le quinte, consapevoli della necessità di una salvaguardia, Hassabis e soprattutto Suleyman stavano istituendo il comitato etico e di sicurezza che Google aveva accettato come condizione per acquisire DeepMind. Google aveva l'obbligo fiduciario verso gli azionisti di aumentare i propri profitti ogni anno, cosa che continuava a fare con successo. Questo garantiva a DeepMind i talenti e le risorse computazionali che servivano, ma rappresentava un'arma a doppio taglio. Una volta creata l'AGI, infatti, Google avrebbe quasi certamente cercato di monetizzarla e controllarla. Non sapevano ancora come, di preciso, ma il comitato avrebbe almeno garantito che la super AI non fosse utilizzata in modo improprio.

Circa un anno dopo l'acquisizione, DeepMind convocò la prima riunione del comitato etico e di sicurezza in una sala conferenze presso la sede di SpaceX, in California. Del comitato facevano parte Hassabis, Suleyman e Legg, così come Elon Musk e Reid Hoffman, il miliardario cofondatore di LinkedIn diventato investitore di venture capital. Tra gli altri presenti, secondo fonti informate, c'erano Larry Page, il dirigente di Google Sundar Pichai, il capo legale di Google Kent Walker, il consigliere postdottorato di Hassabis Peter Dayan e il filosofo della Oxford University Toby Ord.

La riunione andò bene, ma poi i fondatori ricevettero notizie sorprendenti da parte di Google: l'azienda non aveva più intenzione di proseguire con il comitato etico. Suleyman si arrabbiò, dal momento che era stato lui a spingere per la sua istituzione. Parte della spiegazione di Google, in quel momento, era che alcuni membri chiave del comitato avevano conflitti di interesse – Musk, per esempio, stava sostenendo progetti di AI al di fuori di DeepMind – e istituire un comitato non era legalmente fattibile. A molti dei membri del comitato stesso la spiegazione suonò come una scusa. Il sospetto era che, in realtà, a Google non piacesse l'idea di trovarsi alla mercé di un gruppo di persone che potevano sottrarle il controllo di una tecnologia redditizia come quella dell'AI.

Infuriati per quella che sembrava una violazione dell'accordo da parte di Google, Hassabis e Suleyman si lamentarono con i vertici aziendali per lo scioglimento del comitato. Desiderosi di mantenere i fondatori di DeepMind motivati a spingere i confini della ricerca sull'AI, i dirigenti posero loro davanti una ricompensa ancora più

grande. Un alto dirigente di Google li contattò spiegando che forse c'era una struttura migliore per proteggere la loro tecnologia AGI. I fondatori di DeepMind non lo sapevano, all'epoca, ma Google si stava preparando a trasformarsi in un conglomerato denominato Alphabet, cosa che avrebbe permesso alle diverse divisioni aziendali di operare con maggiore indipendenza. Il dirigente disse loro che queste nuove divisioni sarebbero state chiamate «unità autonome». Sarebbe stato come tornare a essere un'azienda indipendente. Avrebbero avuto i loro budget, i loro bilanci, i loro consigli di amministrazione e anche investitori esterni. L'idea sembrava allettante.

In realtà, il vero obiettivo di Google era far salire il suo valore azionario, che ristagnava da tempo. Per anni, gli analisti di Wall Street avevano faticato a valutare l'insieme delle attività di Google al di fuori di YouTube, Android e del suo redditizio motore di ricerca. Tra queste attività, per esempio, c'era un'azienda di termostati intelligenti chiamata Nest, una società di ricerca biotecnologica di nome Calico, una divisione di venture capital e l'X Lab dedicato ai progetti «moonshot». La maggior parte di queste divisioni non generava profitti, ma trasformarle in aziende separate sotto il cappello di un'unica holding avrebbe potuto alleggerire il bilancio complessivo e contribuire a valorizzare ciò che più interessava a Google: la pubblicità. L'attività pubblicitaria di Google rappresentava più del 90 per cento del fatturato annuo. Nonostante la reputazione di azienda tecnologica innovativa e popolata dai migliori ingegneri sulla piazza, la leadership di Google si concentrava ancora principalmente sul più tradizionale dei business: convincere la gente a comprare cose di cui non aveva necessariamente bisogno.

Focalizzati com'erano sulla realizzazione dell'AGI, Hassabis, Legg e Suleyman ebbero a stento modo di ponderare le vere motivazioni di Google, o il fatto che probabilmente l'azienda non aveva alcuna intenzione di concedere loro una reale autonomia, visto che la loro ricerca sull'AI poteva rivelarsi oltremodo utile per far crescere il suo business. L'idea di diventare più indipendenti era semplicemente musica per le loro orecchie, perché significava che Google non avrebbe controllato la loro futura AI, lasciandoli liberi di guidarne gli sviluppi. «Volevamo abbastanza indipendenza per poter affrontare ciò che sarebbe potuto accadere con l'avvento di un'AGI molto potente», ricorda Legg. «Volevamo assicurarci di avere un controllo sufficiente su come si sarebbero sviluppate le cose.»

I fondatori trascorsero l'anno e mezzo successivo a discutere con Page e altri dirigenti su come sarebbe stata la loro esistenza sotto questo nuovo ombrello aziendale e su cosa significasse effettivamente «unità autonoma». Ma poi, quando Google annunciò la ristrutturazione sotto il nome di Alphabet, non confermò alcun piano mirato a dare a DeepMind maggiore autonomia legale. Mentre altre scommesse di Google, come Verily Life Sciences, venivano trasformate in aziende indipendenti, non si registrò alcun progresso in tal senso per DeepMind. Fu come se Google si

fosse dimenticata, ancora una volta, degli impegni presi.

Ma Hassabis non aveva tempo per soffermarsi sul modo in cui Google sembrava continuare a raggirarlo. All'orizzonte, infatti, si profilava una questione ben più preoccupante. A San Francisco, alcuni fondatori di start-up stavano organizzando un altro laboratorio di ricerca con lo stesso obiettivo di DeepMind. L'idea che promuovevano era di costruire un'AGI in modo sicuro e a beneficio dell'umanità. Il sottotesto era un po' una stoccata, poiché implicava che l'altro importante tentativo di realizzare un'AGI – vale a dire, il suo – andasse a beneficio esclusivo di Google. E l'aspetto peggiore era il fatto che questa nuova organizzazione, chiamata OpenAI, era stata fondata da un suo ex investitore: Elon Musk.

6. LA MISSIONE

Era il 2015, e Demis Hassabis costruiva ormai il suo team da cinque anni, raggiungendo sempre nuovi traguardi nel suo percorso lento ma costante verso l'AGI, in un contesto in cui non c'era praticamente nessun altro a tentare la stessa impresa. L'obiettivo di DeepMind era così radicale da permetterle di operare effettivamente come un monopolio. Nessun'altra azienda consolidata al mondo stava cercando di costruire un'AI che potesse superare l'intelligenza umana, perciò Hassabis poteva condurre la ricerca al suo ritmo. Questo, tra l'altro, rafforzava l'idea, tra fondatori e dipendenti, che DeepMind fosse un vero e proprio laboratorio di ricerca orientato alla missione, piuttosto che una semplice azienda. Potevano così venire a patti con il fatto di appartenere a Google, continuando a «risolvere l'intelligenza» (per poi affrontare i più grandi problemi dell'umanità) perché non correvano sulla stessa ruota per criceti a cui la competizione costringeva altre aziende. La loro missione era unica. Ora, l'ipotesi di un rivale nella Silicon Valley stava per cambiare tutto. La marcia verso l'AGI era destinata a trasformarsi in una competizione.

Più cose Hassabis apprendeva sul conto di OpenAI, più la sua rabbia montava. Era stato il primo a intraprendere seriamente il cammino verso la realizzazione di un'AGI e, considerata la marginalità della materia cinque anni prima, per farlo aveva messo a rischio la sua reputazione nel mondo scientifico. Come se ciò non bastasse, questo nuovo competitor, probabilmente, stava sfruttando le sue idee. Sul sito di OpenAI figuravano sette cofondatori: cinque di loro avevano lavorato come consulenti e tirocinanti presso DeepMind per diversi mesi. E questo mandò Hassabis su tutte le furie, dal momento che era stato un libro aperto con il personale di DeepMind riguardo alle diverse strategie da perseguire per raggiungere l'AGI, al modo in cui costruire agenti autonomi o insegnare ai modelli di AI a competere in giochi come scacchi e Go. Ora, cinque scienziati che erano a conoscenza di tutti quei dettagli stavano avviando un'organizzazione concorrente.

Da un punto di vista tecnico, Hassabis non avrebbe dovuto preoccuparsi troppo. C'erano molti altri ricercatori al di fuori di DeepMind che stavano svolgendo un lavoro simile con agenti autonomi, ambienti virtuali e giochi. Uno di quei cinque ex collaboratori era un rinomato scienziato, Ilya Sutskever, specializzato nel *deep learning* e non nella tecnica distintiva di DeepMind, ovvero l'apprendimento per rinforzo. Sutskever era il direttore scientifico di OpenAI e, come i suoi cofondatori, credeva fermamente nelle possibilità dell'AGI.

Tuttavia, Hassabis fremeva per l'audacia di Sam Altman nel reclutare personale che conosceva i segreti di DeepMind, e l'ansia non gli dava tregua. In genere, Hassabis rientrava a casa per cenare con la famiglia, prima di iniziare la seconda parte della sua giornata lavorativa, in cui leggeva articoli di ricerca e inviava e-mail fino alle tre o alle quattro del mattino. In alcune di quelle e-mail o riunioni notturne, Hassabis espresse le proprie preoccupazioni, sostenendo che Altman copiasse la strategia di DeepMind e cercasse di sottrargli i ricercatori.

Hassabis nutriva inoltre qualche perplessità circa la promessa di OpenAI di rendere pubblica la propria tecnologia. Reputava sconsiderato quell'approccio «open». «Pensavo fosse un po' ingenuo credere che l'open-source fosse una panacea», dice oggi. «Man mano che le tecnologie a duplice uso diventano sempre più potenti, che dire dei malintenzionati che potrebbero accedervi per scopi nefasti? […] Non ci sarebbe modo di controllare il fenomeno.» DeepMind pubblicava alcune delle sue ricerche su riviste scientifiche conosciute, ma manteneva il massimo riserbo sui dettagli del suo codice e della sua tecnologia AI. Per esempio, non aveva divulgato i modelli di AI creati per padroneggiare *Breakout*.

Circostanza ancora più umiliante, i leader di DeepMind vennero a sapere che Musk parlava male di Hassabis ai suoi contatti nella Silicon Valley, secondo quanto riferito da persone che lavoravano a DeepMind e OpenAI. Rivolgendosi al nuovo personale di OpenAI, per esempio, il miliardario metteva tutti in guardia circa l'operato di DeepMind in Inghilterra, insinuando che Hassabis fosse un personaggio poco raccomandabile. Gettava ombre sul modo in cui Hassabis aveva progettato *Evil Genius*, un gioco in cui bisognava vestire i panni di un cattivo intento a costruire un'arma apocalittica per dominare il mondo. L'ovvia implicazione era che chiunque creasse giochi del genere, probabilmente, aveva una vena maniacale. Traendo spunto dalla battuta, il team di OpenAI iniziò a creare meme basati sugli screenshot di *Evil Genius* per farli girare su Slack, la piattaforma di messaggistica interna. A un certo punto, secondo un ex elemento dello staff di OpenAI che sentì il commento con le proprie orecchie, Musk arrivò a definire Hassabis l'«Hitler dell'AI».

Qualunque fosse il motivo del suo voltafaccia nei confronti di DeepMind, Musk stava alimentando quella che sarebbe diventata una rivalità feroce. Aveva anche iniziato a sviluppare una visione sull'AI più paranoica e pessimista, in linea con la sua tendenza a estremizzare. Avrebbe potuto, per esempio, battersi semplicemente contro le compagnie petrolifere per far fronte al cambiamento climatico, invece di affannarsi a fare degli esseri umani una specie interplanetaria. Avrebbe potuto acquistare una quota di Twitter, nel momento in cui gli era parsa troppo woke, invece di rilevarla per intero. Forse era l'inclinazione di Musk a prendere misure drastiche, la sua tendenza a esagerare o la convinzione di essere il salvatore dell'umanità, ma un paio d'anni dopo l'investimento in DeepMind, il tycoon stava sprofondando nel dogma dell'*AI doom*, il timore che l'intelligenza artificiale potesse portare a una sorta di apocalisse.

Secondo quanto riportato dal «New York Times», sosteneva lunghe conversazioni

notturne sull'argomento con la moglie, preoccupato dal fatto che il cofondatore di Google Larry Page fosse sulla strada giusta per creare sistemi di AI molto più avanzati dopo aver acquisito DeepMind. Musk e Page erano amici intimi. Frequentavano le stesse cene e le stesse conferenze esclusive, oltre a condividere sogni visionari sul futuro. Stando alla sua biografia scritta dal giornalista di Bloomberg Ashlee Vance, se Musk si trovava a San Francisco e non aveva pensato a un posto per dormire, chiamava Page e gli chiedeva di usare il suo divano. I due giocavano ai videogame e discutevano di aerei futuristici o altre tecnologie. Musk riteneva che Page, sempre più introverso, fosse quasi fin troppo gentile. E questo iniziava a preoccuparlo. Il cofondatore di Google avrebbe potuto creare per errore qualcosa di maligno, come una «flotta di robot potenziati dall'AI in grado di annientare l'umanità». Quella di Musk sembrava una battuta, ma era ciò che effettivamente pensava.

Mesi dopo che Page aveva acquistato DeepMind per 650 milioni di dollari, Musk pubblicò un post riguardante l'AI su un forum online per poi cancellarlo di lì a poco. Nessuno si rendeva conto di quanto rapidamente si stava evolvendo l'AI, diceva il messaggio. «A meno che non abbiate esposizione diretta a gruppi come DeepMind, non potete averne idea.» Musk si diceva inoltre scettico sul fatto che le «principali aziende di AI» potessero impedire che superintelligenze digitali sfuggissero su internet seminando il caos.

Man mano che sprofondava nel tunnel del catastrofismo legato all'AI, Musk cominciò a investire sempre più soldi e tempo nella questione. Donò 10 milioni di dollari al Future of Life Institute, un'organizzazione senza scopo di lucro che promuoveva la ricerca per impedire l'annientamento dell'umanità da parte dell'AI. Poi, insieme a Larry Page, Hassabis e quanti altri erano seriamente impegnati a costruire l'AGI, prese parte a una conferenza organizzata dallo stesso ente a Porto Rico.

Nel corso della conferenza, dopo una cena, Musk e Page ebbero un acceso diverbio. Mentre la discussione si infiammava, sempre più partecipanti si avvicinarono per ascoltare: Page disse che Musk stava diventando fin troppo paranoico riguardo all'AI. Doveva tenere a mente che l'umanità stava evolvendo verso un'utopia digitale in cui le nostre menti sarebbero state a un tempo biologiche e digitali. Continuando a fare tutto quel baccano sull'AI, avrebbe finito per rallentare il percorso verso il futuro.

«Ma come puoi essere così sicuro che una superintelligenza non spazzerà via l'umanità?» chiese Musk.

«Stai facendo dello specismo», pare abba replicato Page, almeno stando al racconto del «New York Times», ergendosi apparentemente a difesa dei postumani del futuro. Concentrandosi sulla catastrofe, Musk stava ignorando le esigenze di tutte quelle future entità destinate a nascere dal silicio.

Da un lato, mentre teneva sotto controllo DeepMind e si calava in una comunità di facoltosi visionari futuristi, Musk diventava sempre più radicalizzato. Dall'altro

lato, però, era in piena crisi da FOMO (da *Fear of missing out*), la paralizzante «paura di rimanere tagliati fuori» che è alla base di alcune delle decisioni d'investimento più importanti nella Silicon Valley. Man mano che l'AI raggiungeva nuove pietre miliari, come la vittoria nella competizione ImageNet del 2012, le grandi aziende tecnologiche cominciavano a prestare maggiore attenzione al settore. Non solo Google aveva acquistato DeepMind, ma Mark Zuckerberg aveva creato una nuova divisione chiamata Facebook AI Research (FAIR), affidandola a uno dei principali esperti mondiali di *deep learning*, Yann LeCun. Probabilmente, fu proprio il desiderio di partecipare a questa nuova corsa all'oro della ricerca a spingere Musk verso un esito controintuitivo rispetto alle sue paure: creare ancora più AI.

In seguito, Musk avrebbe ammesso su Twitter di aver fondato OpenAI per fare da «contrappeso a Google» e perché desiderava che l'AI venisse sviluppata in modo più sicuro. Ma non c'era dubbio che l'AI fosse cruciale per il successo finanziario delle sue aziende, che si trattasse delle capacità di guida autonoma delle auto Tesla, dei sistemi di navigazione dei razzi senza equipaggio di SpaceX o dei modelli alla base della sua futura azienda di interfacce cervello-computer, Neuralink.

Per quanto le visioni apocalittiche rinsaldassero in Musk la convinzione morale di dover raggiungere l'AGI prima di Demis Hassabis, costruire un'AI tanto avanzata quanto quella di Google avrebbe certamente dato impulso ai suoi affari. Si trattava, insomma, di un'impresa redditizia. E solo questo può spiegare perché accettò di lavorare al progetto con uno degli imprenditori più influenti della Silicon Valley: quel Sam Altman che trasformava «milioni» in «miliardi» con le sue presentazioni, che aveva riempito Y Combinator di start-up futuristiche e che riguardo all'AI coltivava ambizioni pari a quelle di Larry Page.

In un'e-mail a Musk del 25 maggio 2015, Altman affermava che qualcuno doveva precedere Google nel realizzare l'AGI, proponendo un progetto strutturato in modo tale che la tecnologia diventasse di dominio pubblico. «Probabilmente vale la pena di parlarne», aveva risposto Musk. Un mese dopo, Altman inviò un'altra e-mail, suggerendo la creazione di un laboratorio per sviluppare «la prima AGI», dando «massima priorità» alla sicurezza. L'AI sarebbe stata di proprietà di un'organizzazione non profit e utilizzata «a beneficio del mondo». Musk rispose: «Sono d'accordo su tutto».

Per Altman, costruire un sistema di AI universale era come prendere tutte le start-up tecnologiche che aveva seguito in Y Combinator e inserirle in un unico, grande coltellino svizzero. Quest'AI avanzata avrebbe potuto avere capacità illimitate. Chi poteva sapere se sarebbero ancora state necessarie aziende o start-up, nel momento in cui una nuova superintelligenza avrebbe potuto generare abbastanza ricchezza da garantire il benessere economico a tutti gli abitanti del pianeta? Se per Hassabis l'AGI avrebbe svelato i misteri della scienza e del divino, Altman la vedeva come la strada verso l'abbondanza materiale a livello planetario. Lui e Musk discussero dell'idea di avviare un laboratorio di ricerca che potesse realizzare questo ideale e fungere da

contrappeso a DeepMind e Google.

Musk e Altman decisero di differenziare ulteriormente la loro nuova organizzazione dalle grandi aziende tecnologiche. Nel suo impegno per costruire un'AI che andasse a beneficio dell'umanità, essa avrebbe collaborato con altri istituti e reso pubblica la sua ricerca. Da qui il nome: OpenAI.

Altman si mise al lavoro per costituire il team di fondatori. Nell'estate del 2015, invitò a cena una dozzina di ricercatori in una sala privata del Rosewood, un hotel di lusso a pochi passi da alcune delle più potenti società di venture capital della Silicon Valley. Tra gli invitati c'erano Ilya Sutskever, lo scienziato che aveva trascorso diversi mesi tra le fila di DeepMind, e Greg Brockman, originario del North Dakota e laureato in matematica a Harvard, già direttore tecnico di Stripe e un vero talento nell'avviare aziende.

Durante la cena, Altman spiegò che l'obiettivo di questa nuova organizzazione sarebbe stato costruire l'AGI e poi distribuirne i benefici al mondo. I convitati trascorsero gran parte della serata a chiedersi se fosse davvero possibile non tanto distribuire all'umanità le ricchezze generate dall'AI, quanto avviare un laboratorio di questo tipo, considerando che le grandi aziende tecnologiche avevano già reclutato la maggior parte dei migliori talenti nel campo. Non era troppo tardi per cercare di assumere i ricercatori più brillanti del settore?

«Sapevamo [inoltre] che le nostre risorse sarebbero state insignificanti rispetto a quelle delle [grandi] aziende tecnologiche», ricordò in seguito Brockman nel podcast di Lex Fridman. E poi, se avessero fondato davvero un'organizzazione simile, come l'avrebbero strutturata per garantire che la sua AI andasse a beneficio del genere umano? «Era chiaro che una tale organizzazione dovesse essere [senza scopo di lucro], senza incentivi che potessero comprometterne la missione.»

A metà strada del viaggio di ritorno in macchina con Altman, Brockman dichiarò che, per quanto il progetto fosse irrealistico, lui ci stava. Dopotutto, quella era la Silicon Valley, il luogo in cui anche le idee più folli trovavano il modo di prosperare.

Essendo lui stesso uno stacanovista, Altman rimase colpito da come Brockman iniziò subito a pianificare tutta la logistica necessaria per avviare OpenAI. Con un tempo medio di risposta alle e-mail pari a cinque minuti, poteva essere altrettanto dedito alla causa quanto lo era lui. «Era dentro da capo a piedi», disse in seguito Altman. Al momento di costruire OpenAI, Brockman si incaricò dell'organizzazione generale.

Brockman si addossò anche il compito di sottrarre un primo gruppo di scienziati di talento ad aziende come Google e Facebook. A tale scopo contattò Yoshua Bengio, un docente della University of Montreal etichettato come uno dei «padrini» del movimento del *deep learning*, non per assumerlo ma perché gli indicasse i ricercatori più promettenti nel campo dell'AI. Bengio stilò un elenco di nomi e glielo inviò.

Assumere queste persone non sarebbe stato affatto semplice. Alcuni di loro guadagnavano stipendi a sette cifre con aziende come Google e Facebook, numeri a

cui Altman e Brockman non potevano nemmeno avvicinarsi. Ciò che avevano dalla loro parte, però, era una missione entusiasmante che puntava a cambiare il mondo e due nomi prestigiosi a capo del progetto. Elon Musk era ormai un tycoon di fama mondiale, e dirigere Y Combinator aveva elevato lo status di Altman nella Silicon Valley a un livello tale che tutti volevano conoscerlo. Per i ricercatori nel campo dell'AI, anche una breve parentesi in questa nuova realtà non profit offriva opportunità prestigiose e un possibile salto di carriera che poteva valere il sacrificio economico.

Numerosi dei principali scienziati nella lista di Brockman accettarono di incontrarlo per discutere del lavoro. Oltre ai grandi nomi e alla visione generale, apprezzavano l'impostazione «open» di questa nuova organizzazione. Avrebbero avuto finalmente la possibilità di *pubblicare* le loro ricerche, invece di lavorare in segreto su qualche prodotto aziendale, e alcuni di loro apprezzavano anche l'idea di contrastare la molla del profitto che motivava gli sforzi di Google e DeepMind verso la realizzazione dell'AI.

Per finalizzare l'accordo, Brockman radunò alcuni scienziati in una cantina. Sutskever sarebbe stato il colpo più grosso di tutti, se avesse accettato di unirsi al team. Il gruppo parlò ancora della creazione di un laboratorio di AI completamente libero dalle pressioni aziendali, tanto da rendere «open-source» la propria ricerca, offrendone i frutti gratuitamente, e di come ciò avrebbe impedito alle grandi aziende tecnologiche come Google e Facebook di esercitare un dominio assoluto sull'AI, man mano che questa diventava sempre più avanzata. Quasi tutti gli scienziati aderirono al progetto, compreso Sutskever, il talentuoso informatico che sorrideva solo di rado. Cresciuto in Russia e in Israele, aveva lavorato con il prestigioso pioniere del *deep learning* Geoffrey Hinton. E ora avrebbe lasciato Google Brain per entrare a far parte di OpenAI.

Con una decina di persone a bordo, nel dicembre del 2015 il team si diresse a Montreal, in Canada, per una conferenza annuale di AI chiamata NIPS (ora NeurIPS), allo scopo di annunciare il nuovo laboratorio di ricerca. Fuori dalla sede dell'evento si accumulava la neve, mentre i membri del team parlavano del laboratorio con gli altri partecipanti alla conferenza. L'annuncio vero e proprio arrivò online. Un sito web, OpenAI.com, fece la sua comparsa con un post di presentazione firmato da Brockman e Sutskever: «Il nostro obiettivo è far progredire l'intelligenza digitale nel modo che più probabilmente avvantaggerà l'umanità nel suo complesso, senza vincoli legati alla necessità di generare ritorni finanziari».

Musk e Altman avrebbero presieduto l'organizzazione, che poteva contare su un impegno di finanziamento imponente, pari a un miliardo di dollari, da parte di Musk, Thiel, Altman, Hoffman e Jessica Livingston, insieme a crediti per il cloud computing da Amazon. Musk stava pianificando di finanziare OpenAI con azioni di Tesla, proprio come si era offerto di fare con DeepMind qualche anno prima.

I numerosi accademici presenti a NeurIPS rimasero basiti nell'apprendere la

notizia. Molti pensavano che costruire l'AGI fosse un sogno irrealizzabile, ma alcuni erano anche invidiosi. Negli ultimi dieci anni, le grandi aziende tecnologiche avevano drenato le università dei loro migliori talenti informatici, e si stava arrivando al punto in cui le menti più brillanti nel campo dell'AI lavoravano per interessi aziendali. Di fatto, ora c'era una catena di montaggio che partiva dalle università di élite e finiva in Google, Facebook e Amazon. Un problema ormai vecchio di anni.

«Non esiste che qualcuno possa dire no a due o tre volte il proprio stipendio», dice Maja Pantic, docente di informatica presso l'Imperial College di Londra passata nel 2018 a dirigere le ricerche del centro di AI di Samsung Electronics per poi trasferirsi in Meta. «È ciò che è successo a me e a tutti i miei colleghi.» Una strada seguita anche da altri luminari. Hinton lavorava ormai per Google; Fei-Fei Li aveva lasciato Stanford per Google; LeCun era entrato in Facebook. Ng aveva lasciato Stanford per Google e poi per Baidu, in Cina. Anche le migliori università, come Stanford, Oxford e il MIT, faticavano a trattenere gli accademici di punta: di fatto, là dove sarebbe dovuta crescere la nuova generazione di educatori, c'era un vuoto. La ricerca sull'AI era sempre più segreta e orientata al guadagno. Ecco perché l'impegno di Musk e Altman nel rendere la loro ricerca accessibile al pubblico rappresentava una ventata d'aria fresca per i ricercatori. Qualcuno, finalmente, stava facendo qualcosa per contrastare la concentrazione dei saperi sull'AI nelle grandi aziende.

La fuga di cervelli dalle università avveniva per due motivi. Il primo e il più ovvio era di carattere economico. Alla University of Toronto, dove il «padrino dell'AI» Geoffrey Hinton aveva insegnato, i docenti di informatica potevano aspettarsi di guadagnare circa 100.000 dollari l'anno. Gli accademici con i salari più alti arrivavano a circa 550.000 dollari l'anno. Quello era il massimo. L'allievo più brillante di Hinton, Sutskever, non aveva nemmeno provato a intraprendere la carriera universitaria. Dopo una parentesi nella start-up di Hinton, era passato direttamente a Google Brain. Quando OpenAI offrì a Sutskever 2 milioni di dollari l'anno, Google Brain triplicò l'offerta, secondo quanto riportato nel già citato *Costruire l'intelligenza*.

Una seconda ragione era legata alla quantità di dati e alla potenza di calcolo necessarie per condurre esperimenti nella ricerca sull'AI. Le università, in genere, dispongono di un numero limitato di GPU, ovvero unità di elaborazione grafica, i potenti semiconduttori prodotti da Nvidia che oggi alimentano la maggior parte dei server utilizzati per addestrare i modelli di AI. Quando lavorava nel mondo accademico, Pantic era riuscita ad acquistare solo sedici GPU per l'intero gruppo di trenta ricercatori. Con così pochi chip, addestrare un modello di AI richiedeva mesi. «Era ridicolo», dice. Non molto tempo dopo essere passata a Samsung, ottenne l'accesso a duemila GPU. Tutta quella potenza di calcolo in più permetteva di addestrare un algoritmo in giorni, anziché in mesi, e la ricerca poteva così procedere molto più rapidamente.

Per quegli scienziati che restavano nel mondo accademico, diventava sempre più

difficile sottrarsi alle Big Tech. Uno studio del 2022 ha rilevato che, nel decennio precedente, il numero di articoli accademici legati alle grandi aziende tecnologiche era più che triplicato, arrivando al 66 per cento. Secondo gli autori dello studio, condotto da ricercatori di diverse università, tra cui la Stanford e lo University College di Dublino, la loro crescente presenza «ricorda da vicino le strategie utilizzate dalle grandi aziende del tabacco». Questo, a sua volta, influenzava il modo in cui le università misuravano il successo della ricerca sull'AI. Secondo Abeba Birhane, che ha guidato lo studio e ora è ricercatrice senior presso la Mozilla Foundation, invece di puntare su valori come il benessere delle persone, la giustizia e l'inclusione, gli accademici tendevano a concentrarsi sulle performance migliori.

Il benessere e l'inclusione non erano concetti vaghi, afferma la Birhane. Erano perfettamente misurabili. «Possono sembrare astratti, ma lo sono anche efficienza e prestazioni», aggiunge. «Le persone hanno trovato modi per misurare equità, privacy e altro ancora.» Ad aggravare ulteriormente la situazione c'era il fatto che, mentre i ricercatori, dalle università alle aziende tecnologiche, puntavano via via a rendere i loro modelli di AI sempre più grandi e capaci, cresceva anche il rischio che questi modelli potessero talvolta produrre esiti razzisti o sessisti, spiega Birhane, citando un altro studio del 2023 di cui è coautrice. «Quello che abbiamo scoperto è che con un dataset più ampio crescono anche i contenuti d'odio.»

Eppure, la capacità computazionale era cruciale per il crescente potere che le grandi aziende tecnologiche stavano accumulando nel campo dell'AI. Google e Meta, l'azienda precedentemente nota come Facebook, disponevano di migliaia di miliardi di dati utili per addestrare i modelli, e gestivano server farm che occupavano centinaia di migliaia di metri quadrati. Un singolo data center attualmente gestito da Google a Dalles, in Oregon, per esempio, è più grande di sei campi da football. La maggior parte delle università può offrire solo una frazione infinitesimale di tali risorse.

Quando l'obiettivo era affinare l'AI, *più* significava anche *meglio*. Nell'avviare la ricerca presso OpenAI, Sutskever e il suo team si concentrarono sulla creazione di modelli di AI il più possibile capaci, non necessariamente equi, giusti o rispettosi della privacy. In termini molto semplici, esisteva una formula per ottenere il risultato. Se addestravi un modello di AI con un numero sempre maggiore di dati, aumentando i parametri a disposizione del modello e la potenza di calcolo impiegata per l'addestramento, il modello di AI diventava più efficiente. Era la stessa straordinaria correlazione osservata dal professor Andrew Ng nei suoi esperimenti a Stanford. Non importava quale fosse l'obiettivo per cui era stato progettato il modello: incrementando tutti i parametri, sarebbe diventato più preciso nella traduzione delle lingue e avrebbe generato testi paragonabili a quelli di un essere umano.

«Se avete un dataset e una rete neurale molto ampi, il successo è garantito», disse Sutskever durante una conferenza sull'AI. Le ultime quattro parole di quest'affermazione divennero il suo motto tra gli scienziati dell'AI, tanto più dopo il

grande lancio di OpenAI, quando il settore rinvigorì il proprio entusiasmo grazie a questa nuova organizzazione senza scopo di lucro, guidata da un brillante scienziato e da alcuni dei magnati più potenti della Silicon Valley.

Non passò molto tempo prima che iniziassero a sorgere alcuni problemi. OpenAI non ottenne subito il miliardo di dollari in finanziamenti annunciato a dicembre da Musk, Thiel e altri. In realtà, di lì a qualche anno, l'organizzazione senza scopo di lucro riuscì a raccogliere solo poco più di 130 milioni di dollari in donazioni effettive, secondo un'inchiesta del sito di notizie tecnologiche TechCrunch, che ha esaminato le dichiarazioni fiscali federali di OpenAI.

OpenAI mancava di fondi e aveva preso una direzione poco chiara. Il team fondatore, composto da trenta ricercatori, iniziò a lavorare nell'appartamento di Brockman nel quartiere Mission di San Francisco, riunendosi intorno al tavolo della cucina o sui divani con i portatili sulle ginocchia. Pochi mesi dopo il lancio, ricevettero la visita di un altro stimato ricercatore di Google Brain, Dario Amodei. Quest'ultimo iniziò a porre domande mirate. Cos'era questa storia di voler realizzare un'AI «friendly» e rilasciarne pubblicamente il codice sorgente? Secondo quanto riportato nel già citato profilo di Altman sul «New Yorker», egli rispose che non avevano intenzione di rilasciare tutto il codice sorgente.

«Ma qual è l'obiettivo?» chiese Amodei.

«È un po' vago», ammise Brockman. Il loro intento era garantire che l'AGI venisse sviluppata nella maniera corretta.

Amodei faceva parte del numero crescente di scienziati che condividevano i timori apocalittici di Musk e Yudkowsky. Lavorava in Google quando, meno di un anno prima, l'azienda era finita nell'occhio del ciclone dopo che si era scoperto come il sistema di riconoscimento visivo di Google Photos classificasse i soggetti di colore come gorilla. Dichiarandosi «inorridita», Google aveva rimosso completamente l'etichetta di gorilla dall'applicazione. «Avere sistemi che falliscono in maniera imprevedibile non è una buona cosa», avrebbe ammesso in un podcast, riferendosi all'incidente.

Ma le preoccupazioni di Amodei non si limitavano alle decisioni razziste e offensive degli algoritmi. Era anche in ansia per come l'apprendimento per rinforzo, la tecnica di AI che DeepMind stava perfezionando, veniva utilizzato per controllare sistemi fisici quali robot, auto a guida autonoma e data center di Google. «Una volta che inizi a interfacciarti direttamente con il mondo e a controllare cose fisiche reali, ritengo che il rischio che qualcosa vada storto […] aumenti in maniera significativa», disse in un'intervista rilasciata nel 2016 al Future of Life Institute di Jaan Tallinn.

La ricerca sui potenziali danni dell'AI lo portò a contemplare scenari sempre più catastrofici, tanto che nel 2023 avrebbe avvertito i media che c'era una probabilità del 25 per cento che un'AI fuori controllo rappresentasse un rischio di estinzione per gli esseri umani. Google Brain non era il posto giusto per trovare soluzioni che riducessero quei rischi. Pochi mesi dopo quella conversazione negli uffici di OpenAI,

Amodei si unì al team.

Per costruire l'AGI, il team fondatore di OpenAI aveva bisogno di attrarre più soldi e talenti; di conseguenza, cercò di concentrarsi su progetti che potessero generare resoconti positivi sui media. I primi ricercatori crearono un computer in grado di battere i migliori campioni umani a *Dota*, un videogioco strategico in 3D, e costruirono anche una mano robotica a cinque dita, alimentata da una rete neurale, capace di risolvere un cubo di Rubik. Questi progetti avevano lo scopo di tenere alto il morale di Elon Musk cercando di sopravanzare il lavoro svolto dall'altra parte dell'Atlantico nelle stanze segrete di DeepMind.

Musk non faceva mistero della sua sfiducia nei confronti di DeepMind. Nel 2017, il personale di OpenAI partecipò a un incontro presso la sede di SpaceX. Musk, che all'inizio visitava gli uffici di OpenAI ogni settimana e poi a intervalli di due settimane, accompagnò i visitatori in giro per la struttura e poi tenne una sessione di domande e risposte con circa quaranta dei suoi nuovi ricercatori nel campo dell'AI. A un certo punto, iniziò a parlare del motivo per cui aveva finanziato OpenAI, e il motivo aveva un nome: Demis Hassabis.

«Ero uno degli investitori di DeepMind e mi preoccupava vedere che Larry [Page] era convinto che Demis lavorasse per lui. In realtà, Demis lavora solo per sé stesso», disse Musk. «E non mi fido di Demis.»

I ricercatori rimasero sbalorditi. A molti di loro sembrò che Musk avesse un conto in sospeso con Hassabis, più che una preoccupazione particolare riguardo alla direzione che stava prendendo l'AI. Quando gli chiesero del suo antagonismo nei confronti di Hassabis, Musk citò i videogiochi che l'imprenditore britannico aveva progettato in passato, incentrati sulla conquista del mondo.

Durante la stessa sessione, Musk riferì di una conversazione avuta con un altro investitore di DeepMind, il quale, parlando di un incontro precedente con Hassabis, aveva detto: «Mi è sembrato di trovarmi in quel punto dei film in cui qualcuno deve alzarsi e sparare al tizio che ha davanti». In altre parole, secondo lui, qualcuno doveva fermare Hassabis prima che realizzasse un'AGI onnipotente.

Per quanto non sembrasse apprezzare Hassabis, Musk continuava a ricordare al personale di OpenAI che DeepMind era in vantaggio, presentando la ricerca dell'azienda britannica come punto di riferimento da raggiungere. Con il passare dei mesi, in Musk crebbe sempre di più la preoccupazione per il fatto che la tecnologia di OpenAI non fosse potente come quella di DeepMind.

Pur di mantenere a bordo il principale finanziatore, Altman e Brockman spinsero alcuni dei ricercatori a emulare il lavoro svolto da DeepMind. I ricercatori del progetto *Dota*, per esempio, non capivano perché dovessero lavorare su una simulazione di videogioco quando l'obiettivo finale era realizzare un'AGI che migliorasse la vita delle persone. Il motivo era che avevano bisogno dei soldi di Musk. «Se non lavoriamo su questo, tra qualche anno o persino l'anno prossimo OpenAI potrebbe cessare di esistere», spiegò loro Brockman.

Sebbene alla fine OpenAI abbia raggiunto una fama mondiale per il lavoro sui chatbot e i modelli di linguaggio di grandi dimensioni, nei primi anni si concentrò su simulazioni multi-agente e apprendimento per rinforzo, ambiti in cui DeepMind era già leader. Ma più inseguivano DeepMind in quei settori, più Altman e il suo team di dirigenti si rendevano conto che questi approcci all'AI non promettevano chissà quale impatto sul mondo reale. Fu allora che OpenAI iniziò a evolvere in un tipo di organizzazione molto diversa da DeepMind. Se quest'ultima, infatti, aveva una cultura gerarchica e accademica che valorizzava gli elementi in possesso di un dottorato, la cultura di OpenAI era più orientata all'ingegneria. Molti dei suoi ricercatori di punta erano programmatori, hacker ed ex fondatori di start-up passati da Y Combinator. Erano più interessati a costruire cose e fare soldi che a raggiungere posizioni di prestigio in seno alla comunità scientifica in virtù di chissà quali scoperte.

Nel frattempo, Musk diventava sempre più impaziente. Si lamentò con Altman del fatto che, pur avendo reclutato una squadra incredibile di scienziati, OpenAI non aveva ancora mostrato qualcosa in cui potesse surclassare DeepMind. Ormai vicina al suo terzo anno di vita, l'organizzazione, secondo Musk, rimaneva ancora troppo indietro rispetto a Google e DeepMind. Così, propose una soluzione rapida: avrebbe preso il controllo di OpenAI e l'avrebbe fusa con Tesla. OpenAI non avrebbe mai colmato il gap con DeepMind senza un cambiamento radicale, spiegò Musk in un'e-mail del dicembre 2018 ad Altman e al suo team (e-mail poi resa pubblica da OpenAI e confermata da qualcuno che aveva letto la versione non censurata). «Purtroppo, il futuro dell'umanità è nelle mani di Demis», aggiunse. In altre parole, se lui non avesse preso in mano la situazione, Hassabis, il cattivo della storia, avrebbe avuto la meglio. Ma Altman e i suoi cofondatori volevano mantenere il controllo, pertanto rifiutarono la proposta di Musk.

Nel febbraio del 2018, in un annuncio pubblico sui nuovi donatori, OpenAI accennò brevemente all'uscita di Musk, presentandola però sotto una luce positiva. Musk, infatti, se ne andava per motivi etici: aveva un insanabile conflitto di interessi nel campo dell'AI. «Elon Musk lascerà il consiglio di OpenAI ma continuerà a contribuire economicamente e a offrire consulenza all'organizzazione», si leggeva sul blog. «Dal momento che Tesla è sempre più focalizzata sull'AI, questa decisione eliminerà un potenziale conflitto futuro per Elon.»

Molti elementi del team di OpenAI sapevano che era una sciocchezza. Sospettavano che, per quanto Musk sbandierasse il proprio interesse nella creazione di un'AI il più possibile sicura, il suo obiettivo era anche quello di essere il primo a costruire l'AI più avanzata. Era già l'uomo più ricco del pianeta e stava acquisendo un'influenza senza precedenti sulle infrastrutture americane: la NASA mandava astronauti nello spazio grazie a SpaceX; Tesla guidava la rivoluzione sugli standard delle auto elettriche; e l'azienda di sua proprietà che si occupava di internet satellitare, Starlink, stava cercando di influenzare l'esito della guerra in Ucraina.

Era evidente che non si poteva fare troppo affidamento su Musk. Aveva promesso di donare un miliardo di dollari a OpenAI nel corso di diversi anni, ma alla fine aveva versato tra i 50 e i 100 milioni di dollari – una cifra irrisoria per l'uomo più ricco tra quanti avevano in qualche modo a cuore la questione dell'AI. Donare quella somma sarebbe stato relativamente facile, soprattutto se avesse finanziato OpenAI con azioni Tesla. Tra il 2015 e il 2023 il valore delle azioni Tesla era aumentato di oltre il 18.000 per cento, perciò OpenAI avrebbe potuto raggiungere l'obiettivo del miliardo di dollari senza troppe difficoltà. A dispetto di tutti i timori professati per il futuro dell'umanità, Musk sembrava molto più interessato al proprio vantaggio sui competitor.

Insieme a Musk, OpenAI perse anche la sua principale fonte di finanziamento. E per Altman, che su quel progetto aveva puntato tutta la sua reputazione, fu un disastro. Alcuni dei migliori scienziati nel campo dell'AI avevano accettato un compenso ridotto pur di lavorare con lui, e le sue pompose promesse di aiutare l'umanità cominciavano a suonare ridicole. Questa nuova era dello sviluppo dell'AI presentava una pura e semplice verità: per avere successo servivano più risorse su ogni fronte, dal denaro con cui pagare i ricercatori ai dati per allenare i modelli, fino ai computer potenti per farli funzionare. Senza Musk, la possibilità di spuntare tutte queste caselle stava rapidamente svanendo.

Altman si trovò di fronte a un bivio cruciale. Lavorando dall'ufficio di OpenAI a San Francisco, rifletteva su come mantenere in vita la non profit con risorse estremamente limitate e sviluppare modelli di AI che, probabilmente, sarebbero stati inferiori a quelli della concorrenza. In alternativa bisognava gettare la spugna e chiudere il progetto. Raccogliere fondi per una non profit era molto più difficile che per una start-up. Altman faticava a convincere persone facoltose a fare donazioni per la causa dell'AGI per pura generosità, senza la possibilità di un ritorno finanziario diretto. Aveva bisogno di decine di milioni di dollari, e Musk era stato il suo ultimo grande benefattore.

C'era un'altra opzione. Forse OpenAI avrebbe potuto offrire ai suoi sostenitori qualche tipo di beneficio economico diretto, oltre all'onore di contribuire alla nascita di un'utopia legata a un'AI che andasse a vantaggio dell'umanità. Sarebbe stato un win-win. Più che una «donazione», i finanziatori avrebbero fatto un «investimento», secondo un linguaggio con cui Altman si sentiva più a suo agio. Tuttavia, vedeva solo pochi potenziali finanziatori a cui potersi rivolgere nella speranza di ottenere sia il denaro sia la potenza di calcolo necessaria per sviluppare l'AGI. E questi erano i colossi tecnologici come Google, Amazon, Facebook e Microsoft. Nessun altro disponeva di miliardi di dollari pronti all'uso o di supercomputer ospitati in edifici grandi quanto svariati campi da football.

Negli ultimi anni, sia OpenAI sia DeepMind avevano cercato di mettere in atto misure atte a impedire che un qualsiasi sistema di AI ultrapotente da loro realizzato trovasse un utilizzo improprio. DeepMind stava tentando di modificare la propria

struttura di governance per evitare che un monopolio motivato dal profitto, come Google, avesse carta bianca nel monetizzare l'AGI. Al contrario, un comitato di esperti avrebbe dovuto tenere sotto controllo la situazione. Altman e Musk avevano fondato OpenAI come una non profit, promettendo di condividere le ricerche e persino i brevetti con altre istituzioni, nel caso in cui fossero giunti alla soglia delle macchine superintelligenti. In quel modo, la priorità sarebbe rimasta il bene del genere umano.

Ora che si trovava costretto a lottare per la sopravvivenza, Altman avrebbe abbattuto alcuni di quei paletti. L'approccio cauto degli esordi si sarebbe fatto più spregiudicato, cambiando radicalmente il campo dell'AI su cui lui e DeepMind stavano lavorando. E così, da impresa ponderata e per lo più accademica si sarebbe trasformata in una specie di Far West. Altman avrebbe sfruttato la sua capacità di imbastire una narrativa convincente per giustificare l'allontanamento dai principi fondatori di OpenAI. Dopotutto, era un imprenditore di start-up, e gli imprenditori di start-up, a volte, dovevano cambiare rotta con manovre drastiche. Così funzionava nella Silicon Valley. Avrebbe dovuto solo *ritoccare* alcuni di quei principi fondatori. Ma giusto un po'.

ATTO II

I LEVIATANI

7. GIOCHI E GIOCHETTI

A pochi passi dalla stazione di King’s Cross, a Londra, dove i turisti affollavano il binario magico che porta a Hogwarts in *Harry Potter*, un altro tipo di magia stava prendendo vita all’interno di una serie di scintillanti grattacieli che si stagliavano nel cielo grigio con facciate di vetro e rivestimenti metallici. Tra l’uno e l’altro si snodava una piacevole passeggiata animata da parecchi frequentatori, alcuni dei quali erano ingegneri e scienziati di DeepMind intenti a recuperare i badge per poi varcare le porte di vetro di un edificio che ufficialmente apparteneva a Google, ma che riservava due interi piani al loro laboratorio segreto di AI.

Nonostante tutti i vantaggi derivanti dall’essere parte di Google, incluse le capsule per il riposo, le stanze per massaggi e la palestra interna, i fondatori di DeepMind stavano ancora cercando di sottrarsi alla presa della loro società madre, Alphabet. Erano passati più di due anni dall’acquisizione, e i dirigenti del colosso tecnologico stavano ponendo Demis Hassabis, Mustafa Suleyman e Shane Legg davanti a una nuova prospettiva: invece che un’«unità autonoma», DeepMind poteva diventare una «società di Alphabet» con bilanci propri.

Lontani dall’implacabile mentalità improntata alla crescita che dominava la Silicon Valley, i fondatori accolsero con fiducia la proposta di Google. Intenzionato a dimostrare che DeepMind poteva camminare con le proprie gambe come impresa, Suleyman si diede da fare per comprovare il valore dei suoi sistemi di AI nel mondo reale. Concentrò una rinnovata attenzione su una divisione che aveva avviato in precedenza, Applied, i cui ricercatori utilizzavano tecniche di apprendimento per rinforzo per affrontare problemi in settori come la sanità, l’energia e la robotica, con l’obiettivo di trasformarli in opportunità d’impresa. Un altro team di circa venti ricercatori, autodefinitosi DeepMind for Google, lavorava su progetti che contribuivano direttamente alle attività della casa madre, per esempio rendendo più efficienti i suggerimenti di YouTube o affinando gli algoritmi per la pubblicità. Stando a una fonte informata sugli accordi, Google accettò di riconoscere a DeepMind il 50 per cento dei ricavi derivanti dal valore aggiunto da queste funzionalità. Circa due terzi dei progetti vennero giudicati utili.

Questo lasciava a centinaia di altri ricercatori di DeepMind la libertà di continuare a studiare nuovi modi per sviluppare l’AGI. Ogni due o tre settimane, i fondatori si ritrovavano in un pub per discutere del lavoro, e le loro conversazioni finivano sempre per toccare punti di tensione ormai familiari. Suleyman voleva risolvere

problemi concreti, ma temeva anche che i loro sforzi potessero finire per dar vita involontariamente a un sistema superintelligente capace di ritorcersi contro l'umanità. Cosa sarebbe successo se l'AI fosse sfuggita dal suo contenitore e avesse cominciato a manipolare la gente? In ufficio, metteva in guardia colleghi e dirigenti circa l'impatto dell'AGI sull'economia, sostenendo che avrebbe potuto causare uno spostamento improvviso di milioni di posti di lavoro e un crollo dei redditi. E se questo avesse provocato una rivolta? «La gente marcerà su King's Cross con i forconi, se non ci preoccupiamo di creare un sistema equo», ripeteva.

Hassabis si sforzava di trovare soluzioni, ma a volte queste sembravano un po' eccentriche. Per esempio, suggerì che, visto che la loro AI diventava più avanzata e potenzialmente pericolosa, DeepMind avrebbe potuto assumere Terence Tao, docente alla University of California di Los Angeles e da molti considerato uno dei più grandi matematici viventi. Ex bambino prodigio entrato all'università all'età di nove anni, secondo la rivista «New Scientist» Tao era una specie di «Mr. Aggiustatutto per ricercatori frustrati».

In alcune interviste, Tao aveva dichiarato che in fondo l'AI non era che matematica avanzata e, con tutta probabilità, il mondo non avrebbe mai avuto una vera AI. Vedeva questa tecnologia in modo meccanicistico e quasi in bianco e nero, proprio come Hassabis. Se l'AI fosse sfuggita al controllo, la matematica avrebbe potuto contenerla. Hassabis non era il solo a pensarla così. Sul forum LessWrong di Yudkowsky, gli utenti avevano avviato un lungo progetto per elaborare strategie tramite cui convincere persone come Tao e altri matematici di spicco a lavorare sull'allineamento dell'AI, ovvero il processo necessario a rendere l'AI più conforme ai valori umani, così da impedirle di andare fuori controllo. La loro ipotesi era che per attirare luminari di quel calibro servissero cifre comprese tra i 5 e i 10 milioni di dollari.

Hassabis immaginava che, una volta avvicinatosi all'AGI, avrebbe smesso di spingere al massimo le prestazioni dei suoi modelli di AI per poi invitare alcune delle menti più brillanti al mondo ad analizzare questi ultimi nei minimi dettagli, così da individuare le migliori strategie matematiche per tenerli sotto controllo. «Forse dovremmo cominciare a lanciare un appello, una sorta di "Avengers Assemble" di matematici e scienziati», dice ancora oggi Hassabis.

Suleyman, tuttavia, non condivideva l'approccio del cofondatore, ritenendolo troppo focalizzato sui numeri e sulla teoria. Credeva che, per essere davvero sicura, l'AI andasse gestita dalle persone, non solo da formule matematiche. Mentre lui e Hassabis discutevano su quale fosse la migliore strategia per contenere l'AI, arrivò un nuovo aggiornamento da parte della leadership di Google riguardo al piano per trasformare DeepMind in una «società di Alphabet». L'idea non poteva funzionare, dissero loro i dirigenti. Creare una nuova realtà non era così semplice perché, man mano che l'AI diventava più preziosa per il business di Google, l'azienda madre aveva sempre più bisogno di DeepMind.

Ai fondatori, questo ennesimo cambio di rotta trasmise una sensazione di déjà-vu. Ma la dirigenza li rassicurò: non c'era motivo di preoccuparsi, perché avrebbero comunque trovato un compromesso. Ora si proponeva una terza opzione: DeepMind avrebbe potuto procedere a una sorta di scorporo parziale e avere il proprio consiglio di amministrazione incaricato di guidare lo sviluppo di un'AI superintelligente, mentre Alphabet avrebbe mantenuto una certa quota dell'azienda. Per dimostrare che intendevano fare sul serio, i vertici di Alphabet misero tutto nero su bianco. La dirigenza firmò un protocollo d'intesa in cui s'impegnava a immettere nelle casse di DeepMind 15 miliardi di dollari in dieci anni, per consentirle di operare in maniera indipendente. Hassabis raccontò a più persone all'interno di DeepMind che la proposta era stata firmata da Sundar Pichai, il dirigente di Google che, di lì a pochi anni, sarebbe diventato CEO di Alphabet. Questo a dimostrazione di come, finalmente, Google sembrasse intenzionata a tenere fede alle promesse.

Il protocollo d'intesa è un documento che delinea i termini e le condizioni di un possibile accordo commerciale. Di solito funge da punto di partenza per ulteriori negoziazioni e non è legalmente vincolante. Tuttavia, un accordo scritto ha un peso maggiore di uno verbale, e i fondatori di DeepMind erano convinti che, questa volta, l'impegno di Google a concedere loro una certa indipendenza fosse concreto. Decisero perciò di dedicarsi completamente alla riorganizzazione di DeepMind in modo da darle una struttura che, al pari di OpenAI, la rendesse più simile a un ente benefico che a un'impresa tradizionale.

Hassabis e Suleyman assunsero banchieri d'investimento per definire i meccanismi finanziari dello scorporo e si affidarono a due studi legali di Londra per redigere i piani giuridici necessari a ristrutturarsi come organizzazione indipendente. Inoltre, chiesero consulenza a un importante avvocato aziendale britannico che aveva gestito operazioni per colossi come Shell, Vodafone e il gigante minerario BHP Billiton.

Pianificarono anche una nuova struttura di leadership: Hassabis, Legg e Suleyman avrebbero fatto parte di un consiglio operativo insieme a Larry Page, CEO di Alphabet, al suo cofondatore, Sergey Brin, al responsabile dei prodotti di Google, Sundar Pichai, e a tre direttori commerciali indipendenti. Le decisioni sarebbero state prese a maggioranza. Elemento fondamentale del piano era l'istituzione di un consiglio fiduciario completamente indipendente, composto da sei garanti incaricati di vigilare sul rispetto della missione sociale ed etica di DeepMind. I nomi di questi garanti e le loro decisioni sarebbero stati pubblici, in modo da garantire la massima trasparenza. Avendo il compito di supervisionare una delle tecnologie più avanzate e potenzialmente pericolose al mondo, questi sei garanti dovevano essere persone di altissima levatura e degne della massima fiducia. Così, DeepMind puntò in alto, tanto da interpellare l'ex presidente Barack Obama, insieme a un ex vicepresidente degli Stati Uniti e a un ex direttore della CIA. Secondo una fonte vicina al progetto, molte tra le personalità contattate accettarono l'incarico.

Dopo una consultazione con esperti legali, DeepMind decise di non intraprendere la stessa strada seguita inizialmente da Sam Altman, ovvero trasformarsi in un'organizzazione senza scopo di lucro. I fondatori concepirono invece una struttura legale completamente nuova, che chiamarono «Global Interest Company» o GIC. L'idea era che DeepMind diventasse un'organizzazione simile a una divisione delle Nazioni Unite, custode trasparente e responsabile dell'AI per il bene dell'umanità. Avrebbe concesso ad Alphabet una licenza esclusiva, in modo tale che eventuali scoperte nel campo dell'AI utili al business della ricerca di Google confluissero nel sistema del colosso tecnologico. Tuttavia, DeepMind avrebbe destinato la maggior parte delle risorse, dei talenti e delle ricerche allo scopo di promuovere la sua missione sociale, lavorando nel campo delle scoperte farmacologiche e dei progressi nel settore sanitario, o nella lotta al cambiamento climatico. All'interno dell'azienda, il progetto veniva chiamato semplicemente GIC.

Eppure, mentre cercava di affrancarsi da Google, DeepMind continuava a rafforzarne il business. Proprio nel periodo in cui prometteva di aiutare DeepMind a separarsi dalla casa madre, Larry Page guardava alla Cina come a una nuova opportunità di espansione. Con il mercato ormai maturo negli Stati Uniti e negli altri paesi occidentali, la Cina rappresentava per Google un'occasione unica. Era il paese più popoloso al mondo, con oltre 650 milioni di utenti internet, quasi il doppio dell'intera popolazione degli Stati Uniti. E considerando che solo la metà dei potenziali utenti era effettivamente online, la Cina rappresentava un mercato vasto e ancora poco sfruttato. La classe media cinese era in crescita, i consumi in aumento e il PIL del paese si attestava intorno agli 11.000 miliardi di dollari, rendendola la seconda economia più grande del pianeta. Era una potenziale miniera d'oro per qualsiasi azienda del web.

Google però non poteva entrare in Cina a suo piacimento. Nel 2010, infatti, aveva lasciato il paese accusando Pechino di aver hackerato la sua proprietà intellettuale e gli account Gmail di attivisti cinesi per i diritti umani. Il governo cinese aveva imposto a Google di censurare le ricerche su argomenti come piazza Tienanmen e altri temi controversi agli occhi del Partito comunista cinese. Poi aveva bloccato l'accesso a Facebook e Twitter, creando quello che sarebbe stato ribattezzato il Grande Firewall. I dirigenti di Google ostentavano comunque una certa sicurezza, ritenendo che si trattasse di una situazione temporanea, perché presto i cittadini cinesi avrebbero preteso i servizi sofisticati e potenti offerti dai giganti web della Silicon Valley.

«La domanda è se, in un arco di tempo sufficientemente lungo, questo tipo di approccio dittatoriale finirà», aveva dichiarato nel 2012 Eric Schmidt, all'epoca presidente di Google, alla rivista «Foreign Policy». «Io credo assolutamente di sì.»

Schmidt si sbagliava. Invece di stagnare, il settore internet cinese ebbe una vera e propria esplosione. Aziende come Meituan, Baidu e Alibaba divennero colossi, mentre ingegneri cinesi che avevano lavorato e avviato aziende nella Silicon Valley

tornavano in patria per costruire i propri giganti tech. Un gran numero di ingegneri provenienti da Microsoft Research Asia assumeva ruoli di leadership in imprese come Alibaba e Tencent. Cinque anni dopo aver lasciato la Cina, Google vedeva il mercato del paese diventare sempre più redditizio ma non aveva una strategia chiara per rientrarvi. Le regole sulla censura, in Cina, non erano cambiate. Ma Google era ansiosa di accedere sia al crescente mercato dei consumatori sia ad alcune delle idee ingegneristiche innovative che stavano fiorendo nel paese. «Dobbiamo capire cosa sta succedendo lì per trarne ispirazione», dichiarò all'epoca Ben Gomes, responsabile della ricerca di Google, alla piattaforma The Intercept. «La Cina ci insegnerà cose che non sappiamo.»

In quel periodo, si verificò un importante cambiamento di leadership ai vertici di Google. Nel 2015, Page e Brin fecero un passo indietro rispetto all'azienda che avevano fondato per perseguire una serie di interessi personali al di fuori di Google, dalla filantropia alle auto volanti fino all'esplorazione spaziale, e nominarono Pichai nuovo CEO. Pichai era il rispettato responsabile dei prodotti che i fondatori di DeepMind avevano intenzione di inserire nel nuovo consiglio operativo, una volta effettuato lo scorporo dall'azienda madre. A differenza di Page, però, non aveva molto tempo (e neppure molto interesse, probabilmente) per concentrarsi sul modo migliore di scorporare una delle acquisizioni più preziose di Google. Lui e Schmidt erano occupati a scovare un qualche sistema creativo per rientrare in Cina. In un certo momento di quell'anno, sembrava che potessero ottenere l'approvazione di Pechino a riportare il loro app store nel paese, ma alla fine non se ne fece nulla.

Poi arrivò un'opportunità di pubbliche relazioni che avrebbe messo DeepMind sotto i riflettori. DeepMind aveva addestrato i suoi modelli di AI con i giochi, e il suo ultimo programma, AlphaGo, era in grado di cimentarsi in partite di Go, un gioco da tavola per due giocatori. Nato in Cina più di due millenni e mezzo anni fa, il Go potrebbe sembrare piuttosto semplice. Si gioca su una griglia di 19×19 con alcune manciate di pietre bianche e nere. I giocatori pongono a turno una pietra su un'intersezione della griglia. L'obiettivo è conquistare territorio sulla tavola circondando i punti vuoti con le proprie pietre e catturando quelle dell'avversario. In realtà, è uno dei giochi più complessi sotto il profilo strategico, dato che il numero di posizioni possibili sulla tavola è dell'ordine di 10^{170}, di gran lunga maggiore rispetto alla stima degli atomi che compongono l'universo osservabile, più vicina a 10^{80}.

Page aveva giocato a Go con il cofondatore di Google Sergey Brin all'epoca in cui stavano avviando l'azienda a Stanford, anni prima, e quando menzionò il suo interesse per il gioco a Hassabis, qualche settimana dopo l'acquisizione, quest'ultimo rispose che il suo team avrebbe potuto costruire un sistema AI in grado di battere un campione umano.

Hassabis non mirava solo a impressionare il suo nuovo capo. Oltre che uno scienziato di successo, era anche un genio del marketing. Sapeva che se AlphaGo fosse riuscito a battere un campione mondiale di Go, come il computer Deep Blue di

IBM aveva battuto Garry Kasparov a scacchi nel 1997, ciò avrebbe segnato un nuovo, entusiasmante traguardo per l'AI e consolidato la credibilità di DeepMind come leader nel campo. Puntati gli occhi sul sudcoreano Lee Sedol, nel marzo del 2016 DeepMind lo sfidò a un incontro di cinque partite a Seul.

Più di duecento milioni di persone si collegarono online e in TV per guardare Lee giocare cinque partite di Go contro il computer di DeepMind. Il ricercatore di DeepMind che gestiva il programma smise di bere ore prima dell'evento per non dover chiedere una pausa per andare in bagno. Durante l'incontro, Hassabis camminava avanti e indietro tra la sala di controllo di AlphaGo e un'area privata di visione. Non riusciva a mangiare. Il suo team aveva insegnato alla rete neurale di AlphaGo trenta milioni di mosse possibili.

Per vincere a Go, i contendenti devono catturare le pietre dell'avversario circondandole completamente, e per farlo occorre una strategia su più livelli: bisogna bilanciare la necessità di attaccare e difendere, valutare obiettivi a lungo termine contro obiettivi a breve termine, nonché prevedere le sequenze di mosse dell'avversario. Questo impone di scegliere con attenzione su quali linee della griglia posizionare le proprie pietre. Le linee più vicine al bordo vengono utilizzate di rado perché non offrono molte possibilità di circondare un avversario per conquistare territorio, perciò la mossa di AlphaGo al trentasettesimo turno della seconda partita contro Lee fu giudicata strana: posizionò infatti una pietra sulla quinta linea della tavola partendo da destra. In genere, le mosse lungo la quinta linea sono considerate meno efficaci perché concedono all'avversario un vantaggio territoriale sulla quarta linea. Insomma, giocare sulla quinta linea era visto come uno spreco. La mossa fu così inattesa e inconsueta che Lee impiegò quindici minuti per riflettere sulla contromossa, arrivando persino a uscire dalla stanza.

«È una mossa sorprendente», disse un commentatore, convinto che l'operatore umano di AlphaGo avesse erroneamente cliccato sulla casella sbagliata della tavola.

Circa cento mosse dopo, però, la strana strategia iniziò ad assumere un senso. Due delle pietre nere di AlphaGo nella parte inferiore sinistra della tavola finirono per sconfinare nella parte opposta, connettendosi perfettamente con la pietra posizionata sulla quinta linea. Dopo altre quattro ore di gioco, Lee si ritirò. Lui e i commentatori avrebbero continuato a definire «fantastica» la trentasettesima mossa. Hassabis, dal canto suo, disse che mostrava bagliori di creatività nell'AI. In totale, AlphaGo vinse quattro delle cinque partite contro Lee.

Questo momento storico per l'AI comportò un'esplosione dell'attenzione mediatica nei confronti di DeepMind, con tanto di premi a un documentario Netflix su AlphaGo. Hassabis era pronto a chiudere in bellezza e a ritirare il programma per passare al progetto successivo.

Ma la dirigenza di Google intravide un'opportunità relativamente al suo disegno di mostrare la potenza tecnologica di Google a Pechino e aprirsi un varco per tornare in Cina. AlphaGo avrebbe potuto rappresentare una nuova forma di diplomazia del

ping pong con la Cina, come lo scambio di giocatori di tennistavolo tra Stati Uniti e Cina nel 1971 che aveva contribuito a scongelare le relazioni diplomatiche in piena guerra fredda. Se la partita in Corea del Sud era stata una trovata pubblicitaria per DeepMind, un secondo incontro organizzato in Cina doveva andare a vantaggio di Google.

Questa voleva che DeepMind facesse confrontare AlphaGo con un giocatore ancora più performante, ovvero il diciannovenne Ke Jie, che allora occupava la posizione numero uno nel ranking mondiale di Go e viveva in Cina. Ke Jie era un giocatore completamente diverso da Lee Sedol: sempre pronto a provocare gli avversari, aspettava solo l'occasione di mettersi in mostra. Ma Google non era meno arrogante nella sua pretesa di potersi guadagnare il ritorno in Cina facendo sfoggio della sua tecnologia.

Hassabis era preoccupato dalla situazione. Se AlphaGo avesse vinto, avrebbe dato l'impressione che la malvagia AI fosse ormai pronta a primeggiare ripetutamente sugli esseri umani. Se avesse perso, tutta l'euforia generatasi a Seul sarebbe svanita. Sembrava una causa persa in ogni caso.

Sapendo però che Google desiderava disperatamente rimettere piede in Cina, Hassabis fece leva sulla sua astuzia strategica per trovare un compromesso con Pichai: avrebbero organizzato un altro incontro, ma questa volta usando una nuova versione di AlphaGo chiamata AlphaGo Master. Invece di funzionare su centinaia di computer diversi, avrebbe operato tramite una sola macchina alimentata da un chip Google. In questo modo, avrebbero potuto presentare l'incontro come un test del loro nuovo sistema AI, piuttosto che come un secondo tentativo di annientare i campioni umani. Se il sistema avesse perso, avrebbero potuto salvare la faccia dicendo che non era comparabile all'AlphaGo originale; se avesse vinto, avrebbero potuto annunciare un nuovo sistema più potente. Google avrebbe potuto promuovere la sua nuova piattaforma di apprendimento automatico chiamata TensorFlow e conquistare qualche grosso cliente in Cina per finanziare la sua piccola ma crescente attività di cloud computing. Pichai fu d'accordo.

L'evento si svolse a Wuzhen, in Cina, nel maggio del 2017, e se i dirigenti di Google avevano passato l'anno precedente a fare pressioni presso i funzionari governativi perché venisse trasmesso via TV e internet in tutta la Cina, alla fine l'incontro rimase oscurato per la maggior parte del paese. Il nuovo AlphaGo vinse tutte e tre le partite in programma contro Ke Jie, ma in Cina quasi nessuno venne a saperlo.

La leadership di Google cercò di mantenere un atteggiamento positivo. Durante un'intervista sul palco dell'evento, Schmidt colse l'occasione per lodare TensorFlow, affermando che le principali aziende internet cinesi come Alibaba, Baidu e Tencent avrebbero dovuto provarlo. «Ne trarrebbero tutte vantaggio», disse. Dietro le quinte, Google era così ansiosa di rientrare nel mercato cinese da rivedere persino alcune delle precedenti resistenze alle richieste di Pechino riguardo alla censura e alla

sorveglianza. Secondo una nota trapelata a The Intercept nel 2018, i dirigenti di Google avevano ordinato ai loro ingegneri di lavorare a un prototipo di motore di ricerca per la Cina (nome in codice: Dragonfly) che bandisse determinati termini dalla ricerca e collegasse le ricerche degli utenti ai loro numeri di telefono. L'azienda stava facendo un passo indietro rispetto ai suoi principi per aiutare un regime oppressivo a sorvegliare i propri cittadini.

Ma la fame di nuovi mercati aveva reso Google completamente incapace di vedere quanto fosse insensato il tentativo di rientrare in Cina. Le aziende tecnologiche cinesi stavano facendo enormi progressi nella ricerca sull'AI e non avevano certo bisogno di TensorFlow o di Google. L'anno prima, il colosso internet Baidu aveva persino assunto Andrew Ng, il professore di Stanford che aveva avviato Google Brain. Il governo cinese aveva capito che i suoi cittadini e il suo fiorente settore tecnologico potevano vivere benissimo senza i servizi del gigante della ricerca online.

Due mesi dopo la partita con Ke Jie, Pechino rivelò il suo ultimo obiettivo a lungo termine per il paese, ovvero quello di diventare leader mondiale nel settore dell'AI, superando gli Stati Uniti già entro il 2030. Il governo avrebbe finanziato una serie di start-up di AI e di progetti futuristici che, presi nel complesso, somigliavano alla sua versione del programma Apollo. Per raggiungere l'obiettivo, non si citarono possibili collaborazioni con Google o con altre aziende tecnologiche della Silicon Valley.

A quel punto, i dirigenti di Google si resero conto che i loro sogni di entrare nel gigantesco mercato internet cinese aumentando i profitti in maniera esponenziale erano irrealistici. Per l'azienda, la delusione fu enorme. Lo stesso Hassabis, con il successo di AlphaGo, si era messo in una posizione scomoda. Creando un'ondata di pubblicità positiva per DeepMind e mostrando la sua AI avanzata, aveva reso il laboratorio ancora più utile agli occhi di Alphabet. Tuttavia, intendeva ancora portare avanti il progetto di scorporo elaborato principalmente con Suleyman.

Era così fiducioso nella riuscita del suo piano che, poche settimane dopo la partita in Cina nel maggio 2017, portò la maggior parte degli oltre trecento membri del team di DeepMind in una zona rurale della Scozia per un ritiro durante il quale, insieme a Suleyman, espose loro il progetto. Nell'hotel e centro congressi preso in affitto, i due annunciarono l'intenzione di trasformare l'azienda in una società di interesse globale a sé stante. Rivelarono quindi che DeepMind sarebbe infine diventata un'organizzazione non profit (con Google come stakeholder) simile ad altre organizzazioni di interesse pubblico come le Nazioni Unite e la Gates Foundation. L'obiettivo era di diventare un'organizzazione orientata al bene comune, spiegarono, e di sovrintendere allo sviluppo dell'AI in modo tale che avesse esiti positivi per l'intero pianeta. Invece di essere un bene finanziario di Google, DeepMind avrebbe stipulato con l'azienda un accordo di licenza esclusivo, continuando però a perseguire la propria missione.

Il team reagì con entusiasmo. Come ricercatore nel campo dell'AI, ecco che

all’improvviso avevi il meglio di entrambi i mondi. Lavoravi per un’azienda tecnologica che offriva un ottimo stipendio e svariati benefit, e intanto pensavi a «risolvere l’intelligenza e poi tutto il resto». I fondatori spiegarono che lo scorporo sarebbe stato finalizzato entro settembre di quell’anno, il 2017.

Hassabis e Suleyman chiesero ai dipendenti di mantenere segreto il progetto GIC, e la richiesta non era poi così insolita visto che la maggior parte dei membri di DeepMind aveva firmato rigidi accordi di non divulgazione che impedivano loro di parlare dei piani e delle tecnologie dell’azienda. In questo caso, però, si chiedeva loro di non parlare dello scorporo nemmeno all’interno dell’azienda. Alcuni ricercatori ricevettero l’istruzione di usare parole in codice, per esempio riferendosi al progetto come *watermelon* («anguria»), e utilizzando app di messaggistica criptate come Signal. Alcuni leader di DeepMind consigliarono persino di non discuterne su dispositivi aziendali o app come Gmail.

I ricercatori di DeepMind ritenevano che la segretezza fosse dovuta alle preoccupazioni riguardo a ciò che Google avrebbe potuto fare con l’AGI. Qualche mese dopo, quando Google venne coinvolta in un progetto militare, quei sospetti acquisirono una certa credibilità. Nel 2017, il dipartimento della Difesa degli Stati Uniti aveva lanciato il cosiddetto Project Maven, con l’obiettivo di una maggiore integrazione dell’AI e dell’apprendimento automatico nelle sue strategie di difesa, fornendo per esempio i suoi droni di *computer vision* per affinare la precisione nell’acquisizione dei bersagli. Quando si unì al progetto, Google si aspettava di guadagnare dalla partnership 250 milioni di dollari l’anno, secondo alcune e-mail fatte trapelare a The Intercept. Tuttavia, di fronte alle massicce proteste interne, Google fu costretta a chiudere il progetto e a rifiutare il rinnovo del contratto con il dipartimento della Difesa, confermando comunque i timori di DeepMind riguardo a un possibile uso improprio della sua AI.

Le operazioni di scorporo però procedevano a rilento. Hassabis e altri dirigenti rassicuravano il personale che alla separazione mancavano sei mesi, per poi ribadire la promessa a distanza di qualche mese. Dopo un po’, gli ingegneri cominciarono a chiedersi se il piano si sarebbe mai concretizzato. E non aiutava il fatto che i contorni del progetto sembrassero piuttosto vaghi. Suleyman, per esempio, diceva al team di voler rendere legalmente vincolanti le nuove regole sui rapporti tra DeepMind e Google, ma né lui né gli altri dirigenti erano in grado di precisare come ciò sarebbe stato possibile nella pratica. Nell’ipotesi che in futuro Google usasse l’AI di DeepMind per scopi militari, DeepMind avrebbe potuto intentare una causa? Non era chiaro. Il team di DeepMind venne incaricato di redigere linee guida che vietassero l’utilizzo della sua AI per violazioni dei diritti umani e «danni di ampia portata». Ma cosa significava nel dettaglio? Nessuno era in grado di dirlo.

Parte del problema stava nel fatto che DeepMind non aveva assunto abbastanza persone per poter rispondere a queste domande. Aveva investito nell’assunzione di scienziati e programmatori per migliorare le prestazioni dei suoi modelli di AI, ma

solo un numero ristretto di dipendenti si occupava di questioni etiche legate alla progettazione di questi modelli. Nel 2020, per esempio, la maggior parte dei circa mille dipendenti di DeepMind era composta da ricercatori scientifici e ingegneri; meno di una dozzina di persone si occupava di etica e solo due di queste conducevano studi accademici sul tema. Quasi nessuno stava analizzando i modi in cui i sistemi di AI potevano generare bias, razzismo o danneggiare i diritti umani. «Non puoi dire di avere un team dedicato all'etica quando in realtà si parla di due persone», disse all'epoca un membro di DeepMind.

In ambito AI, «etica» e «sicurezza» possono riferirsi a obiettivi di ricerca diversi, e negli ultimi anni i loro sostenitori si sono spesso trovati in disaccordo. I ricercatori che si occupano di sicurezza nell'AI tendono a muoversi negli stessi ambienti di Yudkowsky e Jaan Tallinn e puntano ad assicurarsi che un sistema di AGI superintelligente non arrechi danni catastrofici all'umanità, in futuro, utilizzando per esempio la scoperta di nuovi farmaci per creare armi chimiche e sterminare la popolazione, o diffondendo disinformazione su internet fino a destabilizzare completamente la società.

La ricerca etica, invece, si concentra maggiormente sulla definizione di come i sistemi di AI vengano progettati e utilizzati oggi, esaminando i modi in cui la tecnologia potrebbe già adesso arrecare danni alle persone. Questo perché l'algoritmo di Google Photos che ha etichettato i soggetti di colore come «gorilla» non è un caso isolato. Quello dei bias è un problema enorme in ambito AI. Gli algoritmi utilizzati nel sistema giudiziario statunitense hanno etichettato, in maniera non corretta e sproporzionata, gli individui di colore come più propensi a reiterare i reati. Inoltre, alcuni sviluppatori hanno utilizzato gli strumenti di AI per scopi eticamente riprovevoli, come i ricercatori di Stanford che hanno rilasciato un sistema di riconoscimento facciale che affermava di distinguere l'orientamento sessuale delle persone.

Gli sviluppatori dei tre sistemi citati avrebbero dovuto riflettere di più sulla progettazione dei loro modelli, tenendo conto di equità, trasparenza e diritti umani. Tuttavia, si tratta di questioni sfuggenti e di difficile definizione, anche perché in genere non hanno un impatto su quanti gestiscono le aziende di AI (in larga parte, soggetti caucasici di sesso maschile). Oggi, quando i sistemi di AI falliscono, è più probabile che i danni ricadano su persone di colore, donne e altre minoranze.

A lasciare perplessi è il fatto che nel 2017 DeepMind parlava alla stampa e sul proprio sito dell'importanza dell'etica nella sua «missione di risolvere il problema dell'intelligenza per far progredire la scienza in favore di tutta l'umanità.» Alla rivista «Wired», per esempio, aveva anticipato come il piccolo team di ricercatori in campo etico sarebbe cresciuto fino a comprendere venticinque elementi nell'arco di un anno.

In realtà, il team arrivò a soli quindici membri, soprattutto perché, secondo un ex dirigente, i leader di DeepMind erano quasi esclusivamente concentrati sul progetto di scorporo. «Ne parlano tutto il tempo, ma sono solo un pugno di ricercatori e fanno

fatica», dichiarò un altro dipendente, spiegando che il team etico non aveva un gruppo di supporto né risorse adeguate. «Non ha senso. Stiamo parlando di un'azienda multimiliardaria.»

Se DeepMind non metteva in pratica i suoi stessi principi etici, veniva spontaneo chiedersi perché i fondatori fossero così determinati a separarsi da Google. Volevano davvero impedire che la loro tecnologia causasse danni, o intendevano solo soddisfare il desiderio, assai più personale, di mantenere il controllo? Nel quadro dell'accordo di separazione, DeepMind prevedeva di firmare un contratto di licenza esclusiva con Google, ma i fondatori non sembravano in grado di chiarire se l'uso della loro AI per scopi bellici potesse essere vietato, né se tale vincolo fosse legalmente applicabile. Sembravano pieni di ambizioni ma carenti nei dettagli. Alcuni dipendenti iniziarono a chiedersi se la pretesa da parte di Hassabis, Suleyman e Legg di avere la botte piena e la moglie ubriaca – ovvero accettare i finanziamenti di Google per continuare a sviluppare l'AGI ma cercare nel contempo di sottrarle il controllo dell'azienda – non fosse indice di ingenuità.

Così come Google aveva iniziato con il motto *Don't be evil* («Non essere malvagio»), anche i fondatori di DeepMind avevano cominciato la loro avventura sotto Google con le migliori intenzioni. Avevano lasciato sul tavolo di Facebook 150 milioni di dollari pur di mantenere il punto sulla necessità di un comitato etico. Anni dopo, però, sembravano ormai dare priorità alle prestazioni e al prestigio, piuttosto che all'etica e alla sicurezza. Non avevano risposte chiare su come avrebbero contenuto l'AGI, se non assumendo un team di matematici di alto livello come Terence Tao, né su come avrebbero impedito un uso deleterio della tecnologia.

Tutto ciò sollevava una domanda più ampia. Era possibile fare un lavoro significativo su un'AI etica in seno a una grande corporazione? La risposta arrivò dalla stessa Google, ed era un secco no.

8. È MERAVIGLIOSO

Per capire perché sia diventato così tremendamente difficile progettare sistemi di AI etici in Google, o anche solo trasformare idee innovative in prodotti, bisogna fare un passo indietro ed esaminare un paio di dati. Al momento della stesura di questo libro, la società madre di Google, Alphabet Inc., aveva una capitalizzazione di mercato pari a 1.800 miliardi di dollari. Apple era stata la prima azienda statunitense quotata in Borsa a raggiungere una valutazione di 2.000 miliardi di dollari (nel 2020), mentre le valutazioni di mercato di Amazon e Microsoft si aggiravano rispettivamente intorno ai 1.700 e ai 3.000 miliardi di dollari nel 2024. Prima che Apple diventasse la prima società da 1.000 miliardi di dollari, nel 2018, nessuna azienda aveva mai raggiunto simili dimensioni. Eppure, c'è un elemento che accomuna quasi tutte le aziende con i numeri più alti al mondo: sono imprese tecnologiche. In effetti, le aziende che in genere consideriamo colossi hanno solo un quarto delle dimensioni delle controparti nella Silicon Valley. Il gigante petrolifero Exxon Mobil ha una valutazione di mercato di appena 450 miliardi di dollari, mentre Walmart vale 435 miliardi. Se combiniamo la capitalizzazione di mercato dei giganti tecnologici, superiamo il prodotto interno lordo della maggior parte delle nazioni del pianeta, a eccezione degli Stati Uniti e della Cina.

Guardando alla storia, le aziende che un tempo consideravamo giganti impallidiscono se confrontate a quelle di oggi. Al suo apice, prima che venisse smembrata nel 1984, AT&T aveva una capitalizzazione di mercato di circa 60 miliardi di dollari dell'epoca, pari a circa 150 miliardi di dollari odierni. La capitalizzazione massima di General Electric, raggiunta nel 2000, era di circa 600 miliardi di dollari.

Nemmeno il dominio di mercato dei giganti tecnologici ha precedenti. Prima che i regolatori la smantellassero nel 1911, Standard Oil controllava circa il 90 per cento del settore petrolifero negli Stati Uniti. Oggi, Google controlla circa il 92 per cento del mercato globale dei motori di ricerca. Ogni giorno, circa un miliardo di persone in tutto il mondo effettua una ricerca su Google. Più di due miliardi accedono a Facebook. E circa un miliardo e mezzo di persone possiede un iPhone. Nessun governo o impero nella storia ha mai raggiunto un numero così grande di individui in un colpo solo.

Ci sono voluti poco più di due decenni perché queste aziende raggiungessero una portata simile, a partire dal boom e dal susseguente crollo delle dot-com. Come sono diventate così grandi? Hanno acquisito aziende come DeepMind, YouTube e

Instagram, e hanno accumulato una quantità prodigiosa di dati sui consumatori, cosa che ha permesso ad alcune di loro di indirizzarci con pubblicità e suggerimenti in grado di influenzare il comportamento umano su larga scala. Mentre Google raccoglie dati tramite query di ricerca e interazioni su YouTube, Amazon traccia i nostri acquisti e le abitudini di navigazione. La quantità di dati raccolti è difficile da comprendere per le persone comuni: dettagli personali, cronologia di navigazione, dati sulla posizione e, in alcuni casi, persino registrazioni vocali. E queste informazioni non sono soltanto voluminose, ma anche eterogenee, tanto da consentire alle aziende tecnologiche di ottenere un quadro minuzioso del comportamento dei consumatori.

Realtà come Facebook e Google utilizzano questi dati per condurre pubblicità ipermirate, mostrando annunci che catturano l'interesse dei singoli e alimentano sofisticati algoritmi di raccomandazione. Questi software gestiscono i «feed» che gli utenti scorrono ogni giorno, assicurandosi che i contenuti mostrati siano quelli più adatti a mantenerli incollati allo schermo. Le aziende hanno tutto l'interesse a renderci il più possibile dipendenti dalle loro piattaforme, poiché questo genera più introiti pubblicitari. Ma gli effetti negativi sono numerosi. Gli americani sono così dipendenti da Facebook, Instagram e altre app di social media che nel 2023 hanno controllato i loro telefoni in media 144 volte al giorno, secondo uno studio.

Tutta questa «distribuzione personalizzata di contenuti» ha anche esacerbato le divisioni generazionali e politiche tra milioni di persone, poiché i contenuti più coinvolgenti sono spesso quelli che suscitano indignazione. Facebook, per esempio, ha spesso raccomandato i contenuti politici più provocatori nei feed degli utenti durante le elezioni presidenziali statunitensi del 2016, esponendo molte persone a notizie e opinioni che rinforzavano le loro convinzioni, in modo da creare vere e proprie bolle informative. Lo stesso fenomeno ha alimentato il crescente risentimento verso l'immigrazione in Gran Bretagna nei mesi precedenti al referendum britannico sulla Brexit, così come l'ostilità verso il popolo rohingya del Myanmar nel 2017. Secondo un rapporto di Amnesty International, gli algoritmi di Facebook hanno amplificato così tanto la diffusione di contenuti di odio contro i rohingya da favorire la campagna genocida dell'esercito del Myanmar che ha perpetrato uccisioni, torture, stupri e deportazioni di massa nei confronti degli appartenenti a questo gruppo etnico musulmano. Facebook ha poi ammesso a mezzo stampa di non aver fatto abbastanza per prevenire l'incitamento alla violenza contro i rohingya.

Nonostante le divisioni che aveva seminato, il modello di business di Facebook – trattare i suoi miliardi di utenti e i loro dati come prodotto e i suoi inserzionisti come i veri clienti – si è rivelato incredibilmente vincente. Più dati riusciva a raccogliere, più guadagnava dalla pubblicità. Benché questo modello basato sul coinvolgimento abbia avuto effetti tossici sulla società, ha incentivato Facebook a perseguire un unico obiettivo: crescere il più possibile.

L'altro fattore che ha determinato la crescita esponenziale di queste aziende è

stato l'effetto di rete, un fenomeno apparentemente magico che ogni fondatore di start-up sogna di ottenere. L'idea alla base dell'effetto di rete è che quanto più elevato è il numero di utenti e clienti di un'azienda, migliori diventano i suoi algoritmi, cosa che le permette di consolidare ulteriormente la propria presa sul mercato a discapito dei competitor. Nel caso di Facebook, per esempio, le persone hanno iniziato a iscriversi perché tutti gli altri erano già su Facebook, e molti vi sono rimasti per anni – o, almeno, hanno resistito alla tentazione di cancellare il proprio account – per lo stesso motivo. Se siete fan di Apple, saprete quanto sia difficile passare a un altro produttore di dispositivi come Samsung, o far funzionare gli accessori di quest'ultimo con un iPhone. Tutti questi prodotti e servizi interconnessi ostacolano il cambiamento, rafforzando il dominio di Apple.

Non abbiamo un punto di riferimento storico per capire cosa accade quando le aziende diventano così grandi. I valori di capitalizzazione di mercato che Google, Amazon e Microsoft stanno raggiungendo non hanno precedenti. E se da un lato questa crescita porta più ricchezza agli azionisti, compresi i fondi pensionistici, ha anche centralizzato il potere a tal punto che la privacy, l'identità, il dibattito pubblico e, sempre più spesso, le prospettive occupazionali di miliardi di persone dipendono da un pugno di grandi aziende gestite da un numero ristretto di individui incredibilmente ricchi.

Non c'è da meravigliarsi, quindi, se per chi lavora all'interno di un gigante tecnologico e si rende conto che qualcosa non va, lanciare l'allarme può sembrare tanto inutile quanto cercare di far virare il *Titanic* subito prima dell'impatto con l'iceberg. Eppure, tutto ciò non ha impedito a una scienziata di AI, Timnit Gebru, di provarci.

Nel dicembre 2015, alla conferenza NeurIPS in cui Sam Altman ed Elon Musk annunciarono lo sviluppo di un'AI «a beneficio dell'umanità», Gebru si guardò intorno, tra migliaia di altri partecipanti, e rabbrividì. Quasi nessuno, lì dentro, le somigliava. Gebru aveva poco più di trent'anni, era di colore e aveva avuto un'infanzia tutt'altro che convenzionale e priva del sistema di supporto di cui molti dei suoi coetanei avevano beneficiato.

Suo padre, un ingegnere elettronico di origini eritree, era morto quando lei aveva solo cinque anni, e da adolescente era stata costretta a fuggire dall'Etiopia devastata dalla guerra. Non vedendo di buon occhio le sue ambizioni di nuova immigrata, gli insegnanti delle superiori, nel Massachusetts, l'avevano scoraggiata dal frequentare i corsi di Advanced Placement, dicendole che sarebbero stati troppo difficili per lei. Come avrebbe raccontato a «Wired», un insegnante le aveva detto: «Ho incontrato tante persone come te che pensano di poter venire qui da altri paesi e seguire i corsi più difficili». Ma Gebru li aveva frequentati comunque, fino a ottenere una borsa di studio per ingegneria elettronica alla Stanford University.

Alla fine, si era imbattuta nel campo dell'AI e della *computer vision*, un software in grado di «vedere» e analizzare il mondo reale. La tecnologia era affascinante, ma

Gebru aveva subito colto alcuni segnali di allarme. I sistemi di AI stavano assumendo un ruolo sempre più centrale nella vita delle persone: assegnavano un punteggio di credito, concedevano o no un mutuo, segnalavano il volto di qualcuno alla polizia, sostenevano un giudice umano nell'emettere una sentenza penale. Benché potessero sembrare perfetti arbitri neutrali, spesso questi sistemi non lo erano. Se i dati su cui venivano addestrati erano distorti, l'esito logico era inevitabile. E Gebru era dolorosamente consapevole di cosa significassero i pregiudizi.

In un bar di San Francisco, da giovane, era stata attaccata insieme a un'altra donna di colore da alcuni uomini che avevano cercato di strangolarle entrambe. Avevano chiesto aiuto, ma la polizia aveva finito per accusarle di mentire e le aveva trattenute in una cella. Più avanti, mentre scriveva la sua tesi a Stanford, aveva scoperto che in quell'istituto solo un'altra persona di colore aveva conseguito un dottorato in informatica. E dei cinquemila presenti alla più grande conferenza internazionale sull'AI del 2015, quella in cui, appunto, Altman e Musk stavano lanciando OpenAI, solo cinque erano di colore.

Gebru sapeva che non si trattava di casi isolati. Il pregiudizio era sistemico. Decenni dopo l'era dei diritti civili del XX secolo, il razzismo era ancora culturalmente radicato nelle istituzioni e nella psiche del mondo. L'AI rischiava di peggiorare ulteriormente la situazione. Tanto per cominciare, era tipicamente progettata da persone che non avevano sperimentato il razzismo, e questo era uno dei motivi per cui i dati utilizzati per addestrare i modelli di AI spesso non rappresentavano in modo equo le persone appartenenti a gruppi minoritari e le donne.

Gebru ne vedeva le conseguenze nella sua ricerca accademica. Per esempio, si era imbattuta in un'indagine su un software utilizzato dal sistema di giustizia penale degli Stati Uniti chiamato COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), che i giudici e gli ufficiali per la libertà condizionale utilizzavano per prendere decisioni su cauzioni, condanne e libertà vigilata.

COMPAS utilizzava l'apprendimento automatico per assegnare punteggi di rischio agli imputati. Più alto era il punteggio, maggiore era la probabilità che reiterassero il reato. Lo strumento dava punteggi elevati agli imputati neri molto più spesso di quanto non facesse con i bianchi, ma le sue previsioni erano spesso errate. COMPAS era due volte più propenso a sbagliare nelle sue previsioni sul comportamento criminale degli imputati neri rispetto a quanto avveniva con i caucasici, secondo un'indagine del 2016 di ProPublica che ha esaminato settemila punteggi di rischio assegnati a persone arrestate in Florida, per poi verificare se erano state accusate di nuovi crimini nei due anni successivi. Lo strumento era anche più incline a giudicare erroneamente gli imputati bianchi che poi commettevano altri crimini, classificandoli a basso rischio. Il sistema di giustizia penale americano era già sbilanciato nei confronti delle persone nere, e quel pregiudizio sembrava destinato a perpetrarsi con l'uso di strumenti di AI poco trasparenti.

Nella sua tesi di dottorato a Stanford, Gebru aveva portato un altro esempio di

come le autorità potessero usare l'AI secondo modalità inquietanti. Addestrò un modello di visione artificiale per identificare ventidue milioni di automobili mostrate su Google Street View, e poi analizzò cosa potessero rivelare quelle auto sulla demografia di una zona. Quando correlò le automobili con i dati del censimento e della criminalità, scoprì che le zone con più Volkswagen e pick-up tendevano ad avere più residenti bianchi, mentre quelle con Oldsmobile e Buick avevano più residenti neri. E le zone con più furgoni avevano anche più segnalazioni di crimine. Simili correlazioni potevano essere sfruttate. E se la polizia avesse usato quei dati per cercare di prevedere dove potevano verificarsi dei crimini, come in *Minority Report*?

L'idea non era poi così bislacca. Per diversi anni, i distretti di polizia degli Stati Uniti avevano usato i computer per consigliare ai loro agenti quali aree pattugliare, una tecnologia conosciuta come «polizia predittiva». Ma il software veniva addestrato sui dati storici e spesso finiva per indirizzare gli agenti verso comunità di minoranze. Se i dati mostravano che una comunità era sottoposta a una sorveglianza eccessiva, il software continuava a indirizzare la polizia su quella comunità, amplificando un problema già esistente.

L'AI stava diffondendo anche altri stereotipi online, in modo sottile ma insidioso. Sia Google Translate sia Bing Translate a volte rendevano certi mestieri maschili quando venivano tradotti in altre lingue. La frase *o bir mühendis* in turco, lingua che ha pronomi di genere neutro, diventava *he is an engineer* in inglese, mentre *o bir hemşire* diventava *she is a nurse*. Il software operava queste deduzioni grazie a una tecnica detta *word embedding*, che esaminava le parole che tendevano a trovarsi vicino ad altre parole, come *engineer*. Il modello poi determinava quale altra parola si adattasse meglio, come *he*. Google, Facebook, Netflix e Spotify alimentavano tutte le loro raccomandazioni online con la tecnica del *word embedding*, nonostante gli squilibri di genere che introducevano nei propri software.

Era evidente che l'AI avesse problemi che nessuno si era curato di risolvere, perciò, quando Sam Altman annunciò OpenAI, nel 2015, Gebru andò su tutte le furie. Scrisse una lettera aperta denunciando lo spreco di risorse da parte di pochi miliardari egocentrici come Musk e Thiel, che investivano i loro soldi nel tentativo di realizzare un'AI dai poteri divini, e si lamentò del fatto che le uniche preoccupazioni sollevate riguardo alla nuova organizzazione non profit fossero che i suoi ricercatori erano troppo concentrati sul *deep learning*.

«Un magnate della tecnologia bianco, nato e cresciuto in Sudafrica durante l'apartheid, insieme a un gruppo di investitori e ricercatori tutti bianchi e tutti uomini, sta cercando di impedire che l'AI "assuma il controllo del mondo" e l'unico problema che emerge è che "tutti i ricercatori stanno lavorando sul *deep learning*?"» scrisse. «Google ha recentemente rilasciato un algoritmo di visione artificiale che classificava le persone nere come scimmie. COME SCIMMIE. Alcuni cercano di giustificare questo errore affermando che l'algoritmo deve aver considerato il colore come discriminante essenziale nella classificazione degli esseri umani. Se ci fosse stata anche solo una

persona nera nel team, o semplicemente qualcuno che riflettesse sulle questioni razziali, un prodotto che classificava le persone nere come scimmie non sarebbe mai uscito. […] Immaginate un algoritmo che classificasse regolarmente le persone bianche come non umane. Nessuna azienda americana lo definirebbe un sistema pronto per il rilevamento delle persone.»

Uno dei suoi colleghi, però, le consigliò di non pubblicare la lettera. Era troppo esplicita, e probabilmente sarebbe stata identificata. Gebru decise di posticiparne la pubblicazione di qualche anno, ma non poté fare a meno di chiedersi perché alcune delle persone più potenti della Silicon Valley fossero così preoccupate della possibilità di una catastrofe apocalittica legata all'AI quando la stessa AI stava già arrecando danni reali alle persone. Le risposte erano due. La prima era che pochi, se non nessuno, dei leader di OpenAI e DeepMind erano mai stati o rischiavano di diventare vittime di discriminazioni razziali o di genere. La seconda consisteva nel fatto che, paradossalmente, era nell'interesse delle loro aziende sbandierare i pericoli di una superintelligenza onnipotente. Poteva non sembrare sensato lanciare avvertimenti sui pericoli di qualcosa che stai cercando di vendere, ma si trattava di una strategia di marketing brillante. La gente tendeva a preoccuparsi di più per il qui e ora che per il futuro a lungo termine. Se l'AI sembrava destinata ad annientarci in futuro, ciò conferiva anche un fascino seducente alle sue capacità nel presente.

La strategia era anche un modo astuto per distogliere l'attenzione del pubblico dai problemi urgenti e più immediati su cui le aziende avrebbero dovuto intervenire, con la conseguenza, però, di rallentare lo sviluppo e limitare le capacità dei loro modelli di AI. Infatti, per dissuadere i modelli di AI dal prendere decisioni distorte sarebbe stato necessario dedicare più tempo ad analizzare i dati su cui venivano addestrati. Oppure bisognava restringerne il campo di applicazione, compromettendo in tal modo l'obiettivo di dare ai sistemi di AI il potere di generalizzare le loro conoscenze.

Non sarebbe stata la prima volta che grandi aziende distraevano il pubblico mentre i loro affari prosperavano. Nei primi anni Settanta, l'industria della plastica, sostenuta dalle compagnie petrolifere, aveva iniziato a promuovere l'idea del riciclo come soluzione al crescente problema dei rifiuti plastici. Keep America Beautiful, per esempio, un'organizzazione fondata nel 1953 e finanziata in parte da aziende produttrici di bevande e imballaggi, lanciava campagne di servizio pubblico per incoraggiare i consumatori a riciclare. Il suo celebre spot con il pellerossa in lacrime era andato in onda nel 1971 in occasione della Giornata della Terra e spronava la gente a riciclare bottiglie e giornali allo scopo di prevenire l'inquinamento. Chiunque non lo avesse fatto, avrebbe mostrato un disprezzo palese nei confronti dell'ambiente.

Il riciclo non è una cosa negativa di per sé. Tuttavia, promuovendo questa pratica, l'industria poteva sostenere che la plastica non fosse intrinsecamente dannosa, purché venisse riciclata correttamente: in tal modo, spostava la percezione della responsabilità dai produttori ai consumatori. Le aziende della plastica sapevano che riciclare su larga scala era costoso e spesso inefficiente. Un'inchiesta del 2020

condotta da NPR e dalla serie PBS *Frontline* ha scoperto che meno del 10 per cento della plastica viene effettivamente riciclato, nonostante decenni di campagne di sensibilizzazione pubblica.

In ogni caso, queste campagne sono riuscite a impedire che l'attenzione pubblica mettesse in discussione la rapida espansione della produzione di plastica e il danno che stava causando all'ambiente. Il riciclo è diventato parte del discorso pubblico. Testate giornalistiche, consumatori e decisori politici a Washington trascorrevano più tempo a parlare di come riciclare di più che a regolamentare la produzione effettiva di plastica da parte delle aziende.

Proprio come l'industria petrolifera ha sempre distolto l'attenzione del mondo dal suo significativo impatto ambientale, i principali costruttori di AI potevano sfruttare il clamore intorno a un futuro tipo Terminator o Skynet per distrarre dai problemi che gli algoritmi di apprendimento automatico stavano già causando. La responsabilità non ricadeva sui creatori o sull'industria perché agissero subito: era un problema astratto, da affrontare in un secondo momento.

Nel gennaio del 2017, pochi mesi prima che DeepMind cercasse di aiutare Google a rientrare in Cina con AlphaGo, Gebru presentò i risultati della sua tesi a un pubblico di venture capitalist e dirigenti della Silicon Valley. Mentre faceva scorrere le slide, spiegò che i sistemi di AI potevano combinare la capacità di riconoscere le automobili con quella di fare previsioni sui modelli di voto, per esempio, o il reddito familiare.

Uno dei venture capitalist presenti, un investitore di Tesla e amico di Elon Musk di nome Steve Jurvetson, rimase sbalordito, ma non per le ragioni che sperava Gebru. Pensò a quanto potere conferisse questo tipo di dati a Google e alle informazioni che l'azienda avrebbe potuto ottenere riguardo a diversi quartieri o città. Ne rimase così colpito da pubblicare alcune foto della presentazione di Gebru su Facebook.

In quello che rappresentava un costante cortocircuito nel campo dell'AI, alcuni dei presenti coglievano un'opportunità finanziaria mentre altri, come la stessa Gebru, vedevano un pericolo da contenere. Ogni volta che le capacità dell'AI crescevano, emergeva una conseguenza imprevista che spesso causava danni a un gruppo minoritario. I sistemi di riconoscimento facciale erano quasi perfetti nell'individuare i volti degli uomini bianchi, ma cadevano spesso in errore con le donne nere. Uno studio fondamentale del 2018 della ricercatrice del MIT Joy Buolamwini scoprì che i sistemi di riconoscimento facciale di IBM, Microsoft e della cinese Face++ tendevano a classificare erroneamente il genere delle persone dalla pelle più scura e delle donne, cosa di cui lei stessa aveva fatto esperienza. Molti di questi sistemi venivano addestrati su dataset fotografici dominati da uomini caucasici e su immagini raccolte dal web. I database li sovrarappresentavano perché internet rifletteva la demografia delle popolazioni occidentali che avevano maggiore accesso alla rete.

Ma Gebru non aveva intenzione di mollare. Aveva alcune soluzioni. Una di queste prevedeva che i creatori di sistemi di AI seguissero standard più rigorosi

nell’addestrare i loro modelli. Dopo essere entrata in Microsoft, elaborò un insieme di regole chiamato *Datasheets for Datasets*, secondo cui, quando un modello di AI veniva addestrato, i programmatori avrebbero dovuto creare una «scheda tecnica» contenente tutti i dettagli su com’era stato creato, cosa conteneva, come sarebbe stato utilizzato, quali potessero essere i suoi limiti e qualsiasi altra considerazione di carattere etico. Per gli sviluppatori di AI poteva rappresentare un passaggio burocratico in più, ma il procedimento aveva un suo scopo. Se dal modello fossero emersi dei bias, sarebbe stato molto più facile comprenderne il motivo.

Capire perché i sistemi di AI commettono errori è più difficile di quanto si pensi, soprattutto man mano che diventano più sofisticati. Nel 2018, Amazon si rese conto che uno strumento di AI interno, utilizzato per esaminare le candidature di lavoro, continuava a favorire più candidati uomini rispetto alle donne. Il motivo? I creatori del sistema lo avevano addestrato su curriculum inviati all’azienda nei dieci anni precedenti, la maggior parte dei quali proveniva, appunto, da uomini. Il modello aveva imparato che i curriculum con attributi maschili erano più desiderabili. Ma Amazon non riuscì a correggere lo strumento (o non volle farlo) e si limitò a disattivarlo.

Google adottò un approccio altrettanto drastico dopo che il suo strumento di foto aveva etichettato alcuni individui neri come «gorilla», decidendo che l’app non dovesse più riconoscere i gorilla, pur continuando a identificare altri animali. L’errore iniziale, doloroso e offensivo, era nato dal fatto che Google non aveva addestrato il suo strumento con un numero sufficiente di immagini di persone nere e dalla pelle scura e che, probabilmente, non lo aveva testato a sufficienza nemmeno sui propri dipendenti. Alla fine del 2023, l’azienda non era ancora abbastanza sicura di riuscire a correggere il modello di AI; di conseguenza, decise semplicemente di disabilitare la funzionalità.

Alcuni ricercatori di AI sostengono che sia troppo difficile correggere questi bias, perché i modelli di AI moderni sarebbero talmente complessi che nemmeno i loro creatori comprendono appieno il motivo per cui prendono certe decisioni. I modelli di *deep learning*, come le reti neurali, sono costituiti da milioni o miliardi di parametri, noti anche come «pesi», che agiscono come regolatori all’interno di funzioni matematiche complesse tra livelli connessi. Immaginate i livelli di una rete neurale come se fossero una fabbrica con una catena di montaggio in cui ogni operaio sulla linea ha un compito specifico, per esempio verniciare una macchinina o aggiungere le ruote. Alla fine della catena, il risultato è una macchinina completa. Ogni livello in una rete neurale è come una stazione della catena di montaggio: ciascuno di essi apporta una piccola modifica ai dati. Il problema è che, con così tanti microcambiamenti che avvengono in sequenza, è difficile risalire esattamente al contributo di ogni singola stazione sulla catena (o livello nella rete neurale) nel processo di creazione della macchinina (e, allo stesso modo, nel giungere alla decisione di etichettare un imputato di colore come ad alto rischio di recidiva).

Dopo che Google era finita sotto i riflettori per l'incidente dei gorilla, un'altra scienziata informatica, Margaret Mitchell, entrò a far parte del colosso della ricerca per tentare di prevenire simili errori. Nata a Los Angeles e ben nota tra i ricercatori di AI per il suo lavoro sull'equità nell'apprendimento automatico, Mitchell si unì a un piccolo ma crescente movimento che mirava a una maggior cautela circa l'impatto reale dei sistemi di apprendimento automatico. Al pari di Gebru, era preoccupata per gli strani errori che i sistemi di AI continuavano a commettere. Aveva condotto gran parte delle sue ricerche postdottorato in linguistica computazionale e poi nell'ambito della generazione del linguaggio naturale, studiando tutti i modi in cui i computer potevano descrivere oggetti o analizzare emozioni nei testi. In Microsoft, lavorando su un'app per non vedenti, rimase turbata quando il sistema descrisse un individuo caucasico come «persona», mentre un individuo con la pelle scura veniva identificato come «persona nera».

Un'altra volta, mentre conduceva alcuni esperimenti su una rete neurale che descriveva immagini, Mitchell le fornì alcune foto dell'esplosione avvenuta in una fabbrica in Inghilterra. Una di queste era stata scattata dall'alto, da un appartamento vicino, e mostrava colonne di fumo e, in primo piano, una TV sintonizzata su un notiziario che riportava l'incidente. Mitchell rimase sbalordita quando il sistema di AI descrisse l'immagine come «magnifica», «bellissima» e con una «vista straordinaria».

«Il sistema aveva un problema del tipo "è meraviglioso"», racconta Mitchell, riferendosi alla celebre canzone di *The Lego Movie*, film ambientato in un mondo di mattoncini dove tutte le difficoltà della vita vengono allegramente spazzate sotto il tappeto. «Non aveva nessuna nozione della mortalità né del fatto che la morte sia qualcosa di brutto.»

Quello che il sistema aveva effettivamente imparato dalle foto di addestramento era che i tramonti erano belli e che una posizione sopraelevata offriva una vista spettacolare. Fu allora che Mitchell ebbe un'illuminazione: i dati erano tutto. Creando lacune nei dati per addestrare il proprio sistema, lo aveva inconsapevolmente caricato con una serie di bias, compresi quelli che minimizzavano la perdita di vite umane.

Mentre affrontava queste problematiche in Google, Mitchell rilevò un altro aspetto frustrante del lavoro presso una grande azienda tecnologica: si sentiva intrappolata in una burocrazia soffocante, tra riunioni fiume e dirigenti perennemente in ansia per la reputazione dell'impresa.

Nel 2018, Mitchell inviò un'e-mail a Gebru per chiederle di affiancarla in Google. Il settore dell'etica dell'AI era ancora così ristretto che le due si conoscevano già. Gebru sarebbe stata disposta a cogestire il team di ricerca etica sull'AI di Google?

La ricercatrice informatica esitò. Le erano giunte voci di corridoio che descrivevano Google come un ambiente di lavoro tossico, in particolare per le donne e le minoranze. In questo senso il caso di Andy Rubin, dirigente di Google, era esemplare. Rubin era considerato una star all'interno della azienda, avendo cofondato il popolare sistema operativo Android, ma nel 2014 aveva lasciato la società in

sordina in seguito ad accuse di cattiva condotta sessuale. Pochi anni dopo, un'inchiesta del «New York Times» avrebbe rivelato che la dirigenza di Google aveva esaminato le accuse e le aveva ritenute fondate. Eppure, invece di cacciarlo, Google gli aveva riservato un congedo da eroe, con tanto di buonuscita da 90 milioni di dollari.

A ogni modo, non tutto era da buttare. Gebru rimase colpita da come i dipendenti si facevano sentire e reagivano quando l'azienda si comportava in maniera sbagliata. In migliaia avevano organizzato uno sciopero generale per protestare contro il trattamento riservato a Rubin e, pochi mesi prima che lei entrasse in azienda, più di tremila dipendenti avevano firmato una lettera aperta al CEO Sundar Pichai, per chiedere che Google si ritirasse dal Project Maven (cosa che l'azienda fece). Per di più, quelle proteste erano state coordinate da un'esperta di etica dell'AI, una donna di nome Meredith Whittaker, la cui capacità di esporre con chiarezza il problema aveva costretto Google a riconsiderare il programma. Dopotutto, in un posto del genere Gebru avrebbe forse potuto promuovere pratiche più responsabili, come gli standard dettati dai *Datasheets for Datasets*.

Tuttavia, esaminando le dimensioni del suo nuovo team di etica, emergeva chiaramente che le priorità di investimento di giganti tecnologici come Google nell'AI erano le sue capacità. Nonostante l'importanza del lavoro del suo team, questo era composto da appena una manciata di informatici. Nel resto dell'azienda, invece, migliaia di ingegneri e ricercatori continuavano a lavorare per rendere i sistemi di AI più veloci e potenti, creando nuovi standard di capacità che Gebru e Mitchell esaminavano nelle loro conseguenze non volute.

Mitchell si sentiva logorata da Google. Quando, durante le riunioni, metteva in guardia i dirigenti su alcuni dei potenziali problemi che i loro sistemi di AI avrebbero potuto causare, riceveva e-mail dal dipartimento delle Risorse umane che le consigliavano di mostrarsi più collaborativa. Nella Silicon Valley, nel 2020, le donne occupavano appena un quarto dei posti di lavoro nel settore informatico in aziende come Google, Apple e Facebook e guadagnavano ancora solo 86 centesimi per ogni dollaro percepito dagli uomini. Le donne sperimentavano spesso trattamenti diseguali, molestie e discriminazioni nelle assunzioni e nelle promozioni, e la situazione era particolarmente complicata per le donne nere. Molte delle donne che partecipavano alle conferenze o agli eventi informali della Silicon Valley lavoravano nel marketing o nelle pubbliche relazioni, piuttosto che nel reparto ingegneristico o nella ricerca. Erano dunque le donne a occuparsi più spesso di etica dell'AI, sapendo per esperienza diretta cosa significasse la discriminazione. Ma questo le rendeva anche più vulnerabili quando c'era da farsi sentire con autorevolezza nelle discussioni.

Eppure, Mitchell rimase prima sorpresa e poi ammirata da Gebru e dall'audacia con cui non esitava a sfidare l'autorità quando aveva bisogno di risorse o vedeva qualcosa che non andava. Un giorno, sedute nell'ufficio di Gebru nel Building 41 del

campus di Google, si ritrovarono a parlare di un'e-mail inquietante inviata da un dirigente, un messaggio che rifletteva la discriminazione vissuta da entrambe nell'azienda. Mitchell era sul punto di scoppiare in lacrime. Gebru, invece, guardava le cose da una prospettiva diversa.

«Non essere depressa», le disse Gebru. «Arrabbiati.»

Poi tirò a sé il laptop e cominciò a stendere una risposta al manager, leggendo ad alta voce mentre smontava punto per punto, e con precisione chirurgica, ogni questione sollevata dal loro responsabile. In seguito, quando sia Mitchell sia Gebru sarebbero state licenziate da Google, quel dirigente avrebbe preso pubblicamente le loro difese per poi dimettersi poco dopo.

Gebru e Mitchell stavano finalmente per guadagnare la giusta attenzione alla loro causa, anche a costo di essere cacciate dopo aver scatenato quello che sarebbe diventato uno scandalo pubblico. Ma stavano remando contro la priorità centrale di Google. Nel frattempo, il team ben più numeroso di scienziati incaricati di rendere l'AI di Google più intelligente stava per compiere uno dei balzi in avanti più prodigiosi nella storia dell'AI. Un vero e proprio miracolo.

9. IL PARADOSSO DI GOLIA

Nel 2017, Google contava circa 80.000 dipendenti. Non tutti erano ingegneri. C'erano i curatori del Google Doodle che appariva ogni giorno sopra la barra di ricerca. C'erano chiropratici e responsabili dei massaggi in ufficio, «snackologi» con l'incarico di assicurarsi che i tre pasti caldi serviti in mensa tenessero il personale ben nutrito, orticoltori che si prendevano cura delle piante e addetti delle pulizie che lucidavano i tavoli da calcetto.

Il modello di business di Google era una gallina dalle uova d'oro. Quell'anno, il suo settore pubblicitario generò quasi 100 miliardi di dollari – cifra più che raddoppiata di lì al 2024 –, quindi era naturale che gran parte di quei profitti venisse destinata a potenziare il talento interno. Nella Silicon Valley si tendeva a misurare il successo con due parametri: quanto denaro raccoglievi dagli investitori e quante persone assumevi. Un numero spropositato di impiegati rifletteva i sogni di espansione imperiale di CEO come Larry Page e Sergey Brin, anche se non sempre era chiaro cosa facessero molti dei loro manager di medio livello.

L'eccessiva espansione aziendale di Google non era un caso così insolito. All'epoca, Facebook contava circa 40.000 dipendenti e Microsoft 124.000, mentre i fondatori di start-up sognavano di gestire i propri campus aziendali con tanto di palestre e chioschi di gelato gratis. Demis Hassabis era l'eccezione alla regola, forse perché si trovava dall'altra parte dell'oceano. Non voleva che DeepMind venisse risucchiata dalle distrazioni della Silicon Valley o dalle sue ossessioni per i benefit e le dimensioni.

Il problema, in un gigante come Google, era che, se qualcuno al suo interno avesse inventato qualcosa di rivoluzionario, avrebbe faticato comunque a emergere. Il business pubblicitario digitale di Google era sacro. Non era possibile toccare gli algoritmi che lo alimentavano, a meno che non fosse strettamente necessario. Benché la Silicon Valley venisse celebrata come il centro nevralgico dell'innovazione mondiale, le sue aziende più grandi non erano poi così innovative. La homepage di Google era rimasta pressoché invariata nell'ultimo decennio. L'iPhone era sempre il medesimo rettangolo di metallo. E quasi ogni nuova funzione di Facebook era copiata da competitor come Snapchat o TikTok. Una volta che queste aziende avevano raggiunto un fatturato nell'ordine delle decine di miliardi, modificare la formula del loro successo diventava troppo rischioso.

Ecco perché, quando un gruppo di ricercatori di Google fece una delle scoperte

più importanti degli ultimi dieci anni nel campo dell'AI, l'azienda lasciò che languisse inutilizzata. La loro storia, in sintesi, mostrava come la scala monopolistica delle Big Tech ne limitasse la capacità di innovare, costringendole a reagire alle innovazioni altrui copiandole o acquisendole direttamente. Ma questa specifica negligenza si rivelò particolarmente dannosa per Google. Alla fine, OpenAI non solo avrebbe sfruttato a proprio vantaggio la grande invenzione di Google, ma l'avrebbe usata anche per lanciare la prima vera minaccia al gigante della ricerca.

La «T» di ChatGPT sta per *Transformer* («Trasformatore»). Ovviamente questo non ha nulla a che fare con i robot alieni che si trasformano in autoarticolati, ma si riferisce a un sistema che consente alle macchine di generare testi simili a quelli umani. Il transformer è diventato fondamentale per la nuova ondata di AI *generativa* che può produrre testi realistici, immagini, video, sequenze di DNA e molti altri tipi di dati. L'invenzione del trasformatore, nel 2017, ha avuto sul campo dell'AI un impatto simile a quello degli smartphone sui consumatori. Prima degli smartphone, i telefoni cellulari potevano fare ben poco oltre a effettuare chiamate, inviare messaggi e far girare la solita partita di *Snake*. L'arrivo degli smartphone con touchscreen ha permesso agli utenti di navigare su internet, usare il GPS, scattare foto di alta qualità e utilizzare milioni di app.

I trasformatori hanno anche ampliato le possibilità per gli ingegneri di AI, consentendo loro di gestire una quantità di dati molto maggiore e di processare il linguaggio umano assai più rapidamente. Prima dei trasformatori, parlare con un chatbot era come interagire con una macchina ottusa, perché i sistemi precedenti si basavano su set di regole e alberi decisionali. Se chiedevi qualcosa che non era già previsto nel suo programma (circostanza molto probabile), il bot andava in tilt o commetteva errori bizzarri. Questo era il modo in cui erano stati inizialmente progettati assistenti digitali come Siri di Apple, Alexa di Amazon e persino Google Assistant. Trattavano ogni richiesta come una singola istanza isolata, il che non consentiva loro di comprendere il contesto. Non riuscivano a ricordare le domande che avevi posto in precedenza nello stesso modo in cui lo farebbe una persona durante una conversazione. Per esempio:

> «Alexa, com'è il tempo a Indianapolis in questo momento?»
> «In questo momento a Indianapolis ci sono -4 gradi Celsius con cielo nuvoloso.»
> «Quante ore impiegherei per volare da Londra a lì?»
> «Da Londra, per volare fino alla tua posizione attuale, impiegheresti circa quarantacinque minuti.»

Mentre scrivevo queste pagine mi trovavo nel Surrey, che presumibilmente dista circa quarantacinque minuti di volo dall'aeroporto di Heathrow. Poco importava come Alexa avesse elaborato quel piano di volo contorto: il problema era che non riusciva a capire che «lì» si riferiva a Indianapolis, di cui avevo chiesto appena due secondi prima. I sistemi dietro la maggior parte di questi assistenti digitali tradizionali

erano limitati e ancora basati principalmente su parole chiave. Ecco perché continuavano a fornire risposte preconfezionate.

I trasformatori hanno liberato i chatbot da queste limitazioni. All'improvviso, ecco che erano in grado di gestire le sfumature e gli slang. Potevano fare riferimento a ciò che avevi detto poche frasi prima. Potevano affrontare quasi ogni tipo di domanda casuale e fornire una risposta personalizzata. Una sola parola riassumeva il miglioramento: erano più *generali*. E, per molti ricercatori nel campo dell'AI, questo rappresentava un passo verso l'AGI, oltre ad aprire un dibattito su un tema fondamentale: i computer cominciavano davvero a «comprendere» il linguaggio allo stesso modo degli esseri umani o lo stavano semplicemente elaborando attraverso previsioni matematiche?

In un certo senso, è sorprendente che questa invenzione sia venuta fuori da Google. Nonostante il talento e le risorse a disposizione, la complessità della sua struttura e la paura di danneggiare il business pubblicitario costituivano un ostacolo per i dipendenti che cercavano di introdurre innovazioni. Google Brain vantava i ricercatori più avanzati nel campo dell'apprendimento automatico, ma questi ultimi dovevano misurarsi con obiettivi e strategie poco chiari imposti dalla dirigenza. La cultura dell'autocompiacimento derivava in parte dal fatto di poter contare su tanti scienziati di talento, come Geoffrey Hinton. L'asticella era alta, e Google usava già tecniche di AI all'avanguardia come le reti neurali ricorrenti per processare miliardi di parole di testo ogni giorno.

Se foste stati giovani ricercatori come Illia Polosukhin, vi sareste ritrovati accanto alle persone che avevano inventato queste tecniche. Nei primi mesi del 2017, Polosukhin si stava accingendo a lasciare Google ed era disposto a correre qualche rischio. In una delle mense aziendali, due piani sotto l'ufficio di Larry Page, il venticinquenne ucraino stava facendo brainstorming con altri due ricercatori, Ashish Vaswani e Jakob Uszkoreit. Nemmeno i suoi compagni di pranzo amavano seguire le convenzioni adottate dagli altri scienziati dell'azienda. Vaswani desiderava occuparsi di un progetto ambizioso. Uszkoreit, che lavorava per Google da più di dieci anni, guardava con sospetto al modo in cui la struttura degli incentivi di Google Brain si era trasformata in una sorta di istituzione accademica glorificata; dopo aver assunto decine di neolaureati e accademici, si ritrovava circondato da persone il cui primo pensiero era pubblicare articoli o figurare negli atti di una conferenza. Che fine aveva fatto l'obiettivo di creare prodotti fuori dall'ordinario?

Uszkoreit riceveva sguardi ammirati alle feste, non appena rivelava dove lavorava. Ma, ogni volta che precisava di occuparsi di Google Translate, ecco che la gente scoppiava a ridere. Il servizio era macchinoso e spesso impreciso, specialmente con lingue non latine come il cinese. Polosukhin concordò sul fatto che Google Translate facesse abbastanza schifo. Aveva amici, in Cina, che si lamentavano del servizio. E, in quel momento, Uszkoreit si chiese ad alta voce se non ci fosse un modo per migliorarlo. Gli ingegneri di Google erano per lo più convinti di lavorare

già con la tecnologia più avanzata, quindi seguivano il motto: «Se non è rotto, non aggiustarlo». Uszkoreit vedeva le cose in maniera diversa: «Se non è rotto, rompilo».

«E se eliminassimo le reti neurali ricorrenti nel decodificatore della traduzione automatica e usassimo solo il meccanismo di attenzione?» fu la domanda successiva. «Non velocizzerebbe i tempi di inferenza?»

In termini di AI, i ricercatori si stavano chiedendo se fosse possibile sfruttare meglio i potentissimi chip di calcolo. Fino a quel momento, per analizzare le parole, Google aveva utilizzato la tecnica delle reti neurali ricorrenti. Il sistema esaminava ogni parola in una sequenza, come fareste voi leggendo una frase da sinistra a destra. Questo era il metodo all'avanguardia, allora, ma non sfruttava appieno i potenti chip – prodotti da aziende come Nvidia – in grado di gestire più attività contemporaneamente. Il chip del vostro laptop aveva probabilmente quattro «core» per gestire le istruzioni, ma i chip GPU utilizzati nei server per processare i sistemi di AI ne avevano migliaia. Ciò significava che un modello di AI poteva «leggere» molte parole in una frase contemporaneamente, non solo in sequenza. Non sfruttare questi chip era come spegnere una sega elettrica per tagliare il legno manualmente. Immaginate di scollegare una sega elettrica per trascinarne ripetutamente la lama su una tavola di legno. Sarebbe un lavoro lento e faticoso, un inutile spreco del potenziale della macchina. Lo stesso stava accadendo con i sistemi di AI che elaboravano il linguaggio: non stavano usando tutto il potenziale dei chip che li alimentavano.

Ricercatori come Vaswani stavano studiando il concetto di «attenzione» nell'AI, ovvero la capacità di un computer di individuare le informazioni più importanti all'interno di un dataset. Tra un'insalata e un panino, i tre si domandarono se fosse possibile utilizzare quella stessa tecnica per tradurre le parole in modo più veloce e preciso.

Nei mesi successivi, i ricercatori iniziarono a sperimentare. Davanti al silenzioso scetticismo di chiunque li vedesse, Uszkoreit disegnava diagrammi della nuova architettura sulle lavagne dell'ufficio. Ciò su cui stava lavorando il suo team non aveva alcun senso, all'epoca. L'ipotesi era appunto quella di rimuovere l'elemento «ricorrente» delle reti neurali, il che sembrava pura follia. Le diverse architetture che Vaswani stava costruendo, inoltre, non erano ancora significativamente migliori dello *status quo*. Ma quando si sparse la voce del loro progetto, altri ricercatori chiesero di farne parte.

Uno di questi era Noam Shazeer, già una leggenda in Google per aver contribuito a inventare un sistema che aiutava il programma AdSense a stabilire quali annunci mostrare su determinate pagine web. Sempre sorridente e con la voce tonante, era considerato eccentrico e chiacchierava con dirigenti come Sundar Pichai quasi fossero vecchi amici. Shazeer vantava una vasta esperienza con i modelli linguistici di grandi dimensioni, programmi informatici in grado di analizzare e generare testi simili a quelli prodotti dagli esseri umani dopo essere stati addestrati su miliardi di

parole. Poco dopo essersi unito a quel gruppo di ricercatori, Shazeer escogitò alcuni trucchi che aiutarono il nuovo modello a gestire enormi quantità di dati.

«Una volta messe insieme tutte quelle cose, è successo qualcosa di magico», ricorda Uszkoreit. «È stato allora che una serie di altre idee ha preso il volo.»

Ben presto, otto ricercatori lavoravano al progetto, ancora senza nome, programmando in codice e affinando l'architettura di quello che avrebbero battezzato *Transformer*. Il nome faceva riferimento a un sistema capace di trasformare qualsiasi input in qualsiasi output e, benché gli scienziati si concentrassero sulla traduzione linguistica, il loro sistema avrebbe finito per fare molto di più.

Dopo un po', cominciarono a notare alcuni miglioramenti. «Oh, wow, questa è un'altra cosa!» commentò Uszkoreit a un certo punto. Il sistema stava generando strutture frasali lunghe e complesse in tedesco; e, avendo trascorso molti anni in Germania, da bambino, si accorse che era migliore dei soliti contenuti prodotti da Google Translate. Era fluido, leggibile e, cosa più importante, corretto dal punto di vista fattuale. Polosukhin, che parlava francese, notò la stessa cosa.

Llion Jones, un programmatore gallese del team, rimase stupefatto nel constatare che il sistema stava procedendo con la cosiddetta risoluzione delle coreferenze (*coreference resolution*). Questa designa il compito di individuare tutte le espressioni che, all'interno di un testo, si riferiscono alla stessa entità, e aveva rappresentato un grosso ostacolo nel tentativo di far sì che i computer elaborassero il linguaggio in maniera corretta.

Prendiamo per esempio la frase «La gallina non attraversò la strada perché era troppo stanca»: per noi esseri umani è ovvio che «era» si riferisce alla gallina. Nella frase «La gallina non attraversò la strada perché era troppo larga», invece, è altrettanto ovvio che «era» si riferisce alla strada. Fino ad allora, era stato estremamente difficile far comprendere all'AI questo tipo di cambiamento nel contesto, poiché ciò richiedeva un qualche elemento di conoscenza del senso comune, maturata nel tempo attraverso l'esperienza di come funziona il mondo e di come oggetti ed esseri viventi interagiscono.

«È un classico test d'intelligenza in cui l'AI ha sempre fallito», dice Jones. «Non riuscivamo a inserire il senso comune in una rete neurale.» Ma quando alimentarono il trasformatore con le stesse frasi, i ricercatori videro accadere qualcosa d'insolito alla sua unità di attenzione (*attention head*). L'unità di attenzione era come un microrilevatore all'interno del modello che si concentrava su diverse parti dei dati in ingresso. Era la componente che sfruttava la potenza dei chip più avanzati e che permetteva al trasformatore di prestare attenzione a tutte le parole di una frase contemporaneamente, invece che in sequenza.

Quando i ricercatori sostituirono la parola «stanca» con «larga», notarono come l'unità di attenzione spostasse il riferimento dalla gallina alla strada.

«Credo che nessuno avesse mai visto una cosa del genere, prima», ricorda Jones. Gli venne quasi da chiedersi se non stava assistendo a uno sprazzo di intelligenza

reale. «Il fatto che stesse estraendo il senso comune da un testo non strutturato era la prova che stava accadendo qualcosa di più interessante.»

Circa sei mesi dopo quelle prime conversazioni a pranzo, i ricercatori misero per iscritto i risultati ottenuti. Polosukhin aveva già lasciato Google, ma gli altri continuarono a lavorare al progetto, trattenendosi in ufficio fino a mezzanotte. Vaswani, che era l'autore principale, era solito dormire su un divano.

«Ci serve un titolo», disse a un certo punto.

Jones alzò lo sguardo dalla sua scrivania. «Non sono molto bravo con i titoli», rispose. «Ma che ne dici di *Attention Is All You Need*?» Era un pensiero che gli era venuto in mente così, a caso, e Vaswani non fece alcun commento. Anzi, si alzò e se ne andò, ricorda Jones.

Più tardi, però, quel titolo finì sulla prima pagina del loro articolo, sintesi perfetta di ciò che avevano scoperto. Quando si usava un trasformatore, il sistema di AI poteva *prestare attenzione* a grandi quantità di dati contemporaneamente e utilizzarli in svariate maniere.

«Mi piace pensarli come motori di ragionamento», dice Vaswani.

Quei motori di ragionamento avevano il potenziale per rivoluzionare i sistemi di AI, ma Google fu lenta a cogliere l'opportunità. Ci vollero diversi anni, per esempio, prima che l'azienda integrasse i trasformatori in servizi come Google Translate o BERT, un grande modello linguistico sviluppato per migliorare la capacità del motore di ricerca di comprendere le sfumature del linguaggio umano.

Gli inventori del trasformatore non potevano non sentirsi frustrati. Persino una piccola start-up in Germania aveva iniziato a usare il trasformatore per la traduzione automatica ben prima di Google, costringendo il colosso tecnologico alla rincorsa.

Alcuni di loro cercarono di mostrare a Google le possibilità più ampie offerte dal trasformatore. Poco dopo la pubblicazione dell'articolo, Shazeer iniziò a lavorare con un collega per applicare la tecnologia a un nuovo chatbot chiamato Meena. Lo addestrarono su circa quaranta miliardi di parole tratte da conversazioni sui social media e arrivarono a credere che avrebbe rivoluzionato il modo in cui le persone cercavano informazioni sul web e interagivano con i computer. Meena era talmente sofisticato da improvvisare giochi di parole o scambiare battute con un utente umano con la stessa naturalezza con cui poteva sostenere un dibattito filosofico.

Shazeer e il suo collega ne erano entusiasti e cercarono di inviare dettagli del bot a ricercatori esterni, nella speranza di lanciare una demo pubblica e migliorare il rudimentale Google Assistant presente nelle case in forma di altoparlante, per sostituirlo con qualcosa di molto più avanzato. Ma i dirigenti di Google stroncarono sul nascere ogni velleità. Temevano che il bot potesse fare affermazioni sconvenienti, danneggiando la reputazione dell'azienda o, più precisamente, il suo business pubblicitario digitale da 100 miliardi di dollari. Secondo un'inchiesta del «Wall Street Journal», frenarono ogni tentativo di Shazeer di rendere Meena accessibile al pubblico o di integrarlo nei prodotti dell'azienda.

«Google non si muove se non c'è di mezzo un business da un miliardo di dollari», dice Polosukhin. «E avviare un business da un miliardo di dollari è davvero complicato.» Ecco perché così tanti dipendenti di Google avevano lasciato l'azienda per fondare duemila start-up diverse, secondo un'intervista rilasciata da Pichai a Bloomberg nel 2023. Sebbene il dato possa far sembrare Google una fucina di innovazione, in realtà il colosso della ricerca somigliava più a un'enorme piovra che risucchiava ogni innovazione circostante. Molti di quegli imprenditori che avevano avviato nuove aziende finirono per rivenderle a Google o per accettarne gli investimenti. Quando Google non riesce a innovare, di solito compra.

Ci sono due modi di guardare all'approccio cauto di Google verso le nuove tecnologie. Pubblicamente, l'azienda si è sempre mossa con una certa prudenza. E molti ricercatori interni concordano sul fatto che i dirigenti vogliano davvero essere attenti nel distribuire l'AI in maniera tale che non danneggi la società. Negli ultimi anni, Google ha stilato un elenco di principi guida per l'uso dell'AI, copiando in gran parte regole già elaborate da DeepMind. Nel 2018, il capo del suo dipartimento legale, Kent Walker, annunciò che l'azienda avrebbe smesso di vendere tecnologie di riconoscimento facciale a causa del potenziale rischio di abusi. Più in generale, Google sottopone i suoi algoritmi a un rigoroso processo di revisione interna, che talvolta include anche l'intervento di esperti esterni per valutare eventuali problemi etici.

Tuttavia, l'azienda continuava a prendere decisioni eticamente discutibili. Nel maggio di quello stesso anno, Pichai presentò una nuova funzionalità dell'assistente vocale chiamata Duplex: un'AI in grado di telefonare a un ristorante per prenotare un tavolo, utilizzando un intercalare infarcito di «ehm» e «uh» al punto da sembrare umana in modo inquietante. Pichai concluse la dimostrazione tra applausi scroscianti, ma il servizio non rivelava di essere una macchina. Google fu subito criticata per aver ingannato le persone all'altro capo del telefono.

L'approccio prudente di Google era in gran parte il risultato della sua lentezza burocratica. L'altra faccia della medaglia, per una delle aziende più grandi di sempre, con un monopolio quasi assoluto sul mercato della ricerca online, è che tutto si muove a passo di lumaca. Il timore costante di una reazione pubblica negativa o di controlli normativi paralizza l'innovazione. La principale priorità diventa sostenere la crescita e preservare la posizione dominante. Google è stata così ossessionata dal mantenere la sua presa sul mercato della ricerca che nel 2021 ha pagato oltre 26,3 miliardi di dollari ad Apple, Samsung e altri – più di *un terzo* del suo utile netto di quell'anno – solo per preinstallare il suo motore di ricerca sui loro dispositivi, com'è emerso in una recente causa antitrust intentata dal dipartimento di Giustizia degli Stati Uniti.

Le dimensioni colossali dell'azienda e la sua mania di crescita costringevano i ricercatori e gli ingegneri a superare numerosi livelli di gestione anche solo per far approvare delle minuzie. E con una concorrenza praticamente inesistente – dato che

Google controllava circa il 90 per cento di tutte le ricerche online nel mondo – non c'era alcuna urgenza di innovare.

A un certo punto, mentre il gruppo dei trasformatori affinava le proprie ricerche, Shazeer si ritrovò a chiacchierare direttamente con Pichai accanto a una delle tante macchinette del caffè sparse per l'azienda. I lunghi trascorsi come esperto di AI all'interno di Google gli avevano permesso di instaurare rapporti personali con alcuni dei massimi dirigenti. Secondo uno dei coautori dell'articolo sui trasformatori, Łukasz Kaiser, presente alla conversazione, Shazeer disse con la massima sicumera: «Questa cosa rimpiazzerà completamente Google».

«Aveva già questa convinzione, ovvero che avrebbe sostituito tutto», ricorda Kaiser. Shazeer diceva più o meno le stesse cose anche ai colleghi e aveva promosso il potenziale rivoluzionario del trasformatore in un promemoria interno indirizzato alla dirigenza di Google; dunque, non stava affatto scherzando. Il trasformatore permetteva ai computer di generare non solo testi, ma risposte a ogni tipo di domanda. Se i consumatori avessero iniziato a usare sempre più di frequente uno strumento simile, avrebbero potuto finire per trascurare progressivamente Google.

Pichai sembrò liquidare il commento, catalogando Shazeer come uno dei ricercatori più eccentrici di Google e limitandosi a dire: «Va bene, approfondite la questione». In preda alla frustrazione, Shazeer lasciò Google nel 2021 per proseguire autonomamente la sua ricerca sui modelli linguistici di grandi dimensioni, cofondando un'azienda di chatbot chiamata Character.ai. A quel punto, l'articolo *Attention Is All You Need* era diventato uno dei lavori di ricerca più influenti di tutti i tempi nel campo dell'AI. Di norma, un articolo scientifico sull'AI poteva ricevere qualche decina di citazioni, se i suoi autori erano fortunati. Ma il paper sui trasformatori ebbe un impatto tale, tra gli scienziati, da guadagnarsene più di ottantamila.

Non c'era nulla di insolito nel fatto che Google condividesse con il mondo alcune delle meccaniche di base di un'invenzione: era così che operavano spesso le aziende tecnologiche. Quando rendevano «open-source» nuove tecniche, ricevevano feedback dalla comunità scientifica, accrescendo così la loro reputazione tra i migliori ingegneri e facilitandone le assunzioni. Ma Google sottovalutò il costo che avrebbe pagato questa volta. Nessuno degli otto ricercatori che hanno inventato il trasformatore lavora più per l'azienda. La maggior parte di loro ha fondato una propria società di AI, e al momento della stesura di questo testo il loro valore complessivo superava i 4 miliardi di dollari. Character.ai, da sola, valeva un miliardo di dollari ed era destinata a diventare uno dei siti di chatbot più popolari al mondo. Grazie a quell'innovazione che Google non ha saputo sfruttare in maniera adeguata, Shazeer è ora proiettato verso la stratosfera: «La ricerca online è una tecnologia da mille miliardi di dollari, ma mille miliardi non sono *cool*», dice oggi dal suo ufficio a Menlo Park, in California. «Sai cos'è *cool*? Un milione di miliardi di dollari. Questa è una tecnologia da un milione di miliardi di dollari, perché se la ricerca aveva lo scopo

di rendere universalmente accessibili le informazioni, l'AI ha lo scopo di rendere universalmente accessibile *l'intelligenza* e, in tal modo, rendere tutti enormemente più produttivi.»

Dopo la partenza di Shazeer, Google tenne per sé la sua ricerca su Meena, che in seguito ribattezzò Language Model for Dialogue Applications, o LaMDA. I suoi scienziati continuarono a lavorare sul modello, addestrandolo e perfezionandolo con l'aiuto di collaboratori esterni fino a renderlo fluido e, con loro sorpresa, incredibilmente simile a un essere umano.

Per quanto questi progressi fossero entusiasmanti, Google voleva tenere tutto confinato nella sua bolla interna: LaMDA era probabilmente il chatbot più avanzato al mondo, ma solo poche persone all'interno di Google potevano utilizzarlo. L'azienda era riluttante a rilasciare qualsiasi nuova tecnologia potesse finire per minacciare il successo della sua attività principale, la ricerca online. I dirigenti e il team di pubbliche relazioni la presentarono come una strategia improntata alla prudenza ma, più di ogni altra cosa, l'azienda era intenzionata a mantenere la propria reputazione e lo *status quo*.

Ben presto, Google avrebbe vissuto quello che Ashish Vaswani definisce un «momento biblico». Mentre il colosso di Mountain View continuava a stampare denaro grazie al business pubblicitario, OpenAI sembrava in procinto di compiere un passo monumentale verso l'AGI e, a differenza di Google, non stava mantenendo il segreto.

ATTO III

I CONTI

10. LE DIMENSIONI CONTANO

Uscendo dal quartier generale di Google, nella soleggiata Mountain View, in California, e guidando verso nord per circa un'ora, alla fine arrivereste a San Francisco e, una volta scesi dall'auto, sareste attraversati da un brivido. Qui, in genere, la temperatura è di diversi gradi più bassa, con nuvole grigie a incombere basse nel cielo. Mentre nella città natale di Google si gira in maglietta, nel microclima urbano di OpenAI serve una giacca. Un'altra grande differenza, all'epoca, era questa: i ricercatori di OpenAI erano entusiasti della tecnologia dei trasformatori, che invece la dirigenza di Google intendeva tenere chiusa in un armadio metaforico. Ai ricercatori con base nella fredda San Francisco stava per venire un'idea.

Le due dozzine circa di ricercatori del laboratorio non profit erano ancora impegnate a cercare di eguagliare il successo di DeepMind e non vedevano l'ora di realizzare un grande progresso nel campo dell'AI. Avevano assistito alla vittoria di AlphaGo contro i migliori giocatori di Go a livello globale e ora stavano addestrando i loro agenti AI a giocare a *Dota 2*, un complesso videogioco strategico simile a *World of Warcraft*. Se un agente AI fosse riuscito a guidare un elfo in un mondo fantastico, forse avrebbe potuto catturare la natura caotica e incessante del mondo reale meglio di quanto avesse fatto AlphaGo di DeepMind. A prima vista, sembrava ben più impressionante che muovere un mucchio di pietre bianche e nere su una tavola da gioco.

Nel frattempo, tra Sam Altman e Demis Hassabis era in corso una sorta di mini guerra fredda, e il conviviale membro del comitato di OpenAI Reid Hoffman stava cercando un modo per convincerli a «fumare il calumet della pace», secondo quanto riferito da una fonte che ha sentito il commento in prima persona. Nel 2017, sia Altman sia Hassabis parteciparono a una conferenza sulla sicurezza dell'AI, organizzata in California dal Future of Life Institute. Hoffman era presente e, dopo l'evento, tentò di portare a cena il guru delle start-up americane e il neuroscienziato britannico. Altman non gradì l'idea, sostenendo che Hassabis fosse poco collaborativo e apparentemente indifferente ai rischi esistenziali dell'AI che lui, invece, stava cercando di prevenire. Così, Hoffman invitò Mustafa Suleyman al suo posto. I due concordarono sulla volontà di rendere il mondo un posto migliore, e per un po' sembrò che le rispettive organizzazioni potessero trovare un punto d'intesa.

Dietro le quinte, però, Altman e Hassabis si contendevano i migliori ingegneri. Grazie al suo finanziatore Big Tech, Hassabis si trovava in una posizione di

vantaggio e poteva offrire ai ricercatori dell'AI compensi ben più elevati di quelli che poteva permettersi Altman, oltre a stock option di Google. Hassabis non disdegnava di inviare qualche e-mail alla dirigenza di OpenAI per ricordarle che non aveva rivali in quella corsa ai talenti. I manager di OpenAI mostravano quelle e-mail agli ingegneri che stavano cercando di reclutare. «Se era convinto che non avessimo possibilità di successo, perché mai inviarci queste e-mail?» ricorda un ex dipendente di OpenAI.

Forse lo faceva perché lo stesso Altman, a sua volta, aveva l'abitudine di contattare direttamente gli ingegneri di DeepMind per vedere se fossero disposti a cambiare squadra. In ogni caso, stando a un altro ex dipendente, adottava un approccio prudente e mirato al reclutamento, dedicandovi circa il 30 per cento del suo tempo e parlando a lungo con ogni candidato. «Siamo andati a casa sua e abbiamo camminato un'ora intera per Russian Hill, a San Francisco», racconta un ex membro del team ricordando il suo colloquio con Altman. Per chi entrava nella squadra, poi, Altman rimaneva largamente accessibile, seduto con il suo laptop nell'ufficio open space dell'azienda. «Chiunque poteva scrivergli su Slack e parlargli», ricordano in molti. «Non c'era nulla di male.» Nella struttura più gerarchica di DeepMind, invece, Hassabis tendeva a rintanarsi nel suo ufficio o in una sala riunioni ed era più difficile da avvicinare. Per ottenere un incontro con lui, bisognava passare attraverso altri manager e intermediari.

OpenAI stava per distinguersi da DeepMind in un altro modo. Ilya Sutskever, il suo scienziato di punta, non riusciva a smettere di pensare a cosa potesse fare il trasformatore con il linguaggio. Google lo stava usando per comprendere meglio i testi. E se OpenAI lo avesse usato per *generarli*, quei testi? Sutskever parlò con un giovane ricercatore di OpenAI, Alec Radford, che stava sperimentando con i modelli linguistici di grandi dimensioni. Benché OpenAI sia oggi principalmente conosciuta per ChatGPT, nel 2017 stava ancora esplorando diversi scenari, cercando di capire cosa potesse funzionare, e Radford era uno dei pochi a concentrarsi sulla tecnologia che alimentava i chatbot.

I modelli linguistici di grandi dimensioni, all'epoca, erano ancora considerati una barzelletta. Le loro risposte erano per lo più preimpostate e spesso commettevano errori madornali. Radford, che portava gli occhiali e una zazzera rossiccia da liceale, era impaziente di superare tutti i precedenti tentativi accademici di rendere i computer più capaci di parlare e ascoltare, ma era anche un ingegnere e puntava a progressi più rapidi. Da almeno sei mesi si scontrava con ostacoli invalicabili nei suoi esperimenti, passando settimane su un progetto per poi abbandonarlo in favore del successivo. Aveva addestrato un modello linguistico su due miliardi di commenti prelevati dal forum online Reddit, ma era stata un'operazione infruttuosa.

In un primo momento, con l'invenzione del trasformatore da parte di Google Radford accusò il colpo. Era evidente che un'azienda di quelle dimensioni avesse più competenze nel campo dell'AI. Dopo un po', tuttavia, divenne altrettanto palese che Google non avesse grandi piani per la sua nuova invenzione, e Radford e Sutskever si

resero conto che avrebbero potuto sfruttarne l'architettura a vantaggio di OpenAI. Dovevano solo darle un'impronta personale. Il modello trasformatore che alimentava Google Translate usava una struttura composta da un encoder e un decoder per elaborare le parole. L'encoder elaborava la frase in ingresso, magari in inglese, e il decoder generava l'output, per esempio una frase in francese.

L'idea era paragonabile a una conversazione tra due robot. Il primo, l'encoder, ti ascoltava e prendeva appunti, poi li passava al secondo, il decoder, che li leggeva e ti rispondeva. Radford e Sutskever capirono di poter eliminare il primo robot e lasciare, invece, che fosse il decoder ad ascoltare e rispondere. I primi test mostrarono che l'idea funzionava anche nella pratica, il che significava che avrebbero potuto costruire un modello linguistico più snello, più veloce da correggere e alimentare. E il fatto che fosse basato solo sul decoder si rivelò rivoluzionario. Combinando la capacità del modello di «comprendere» e parlare in un unico processo fluido, si poteva arrivare a generare testi molto più simili a quelli umani.

Il passo successivo sarebbe stato aumentare enormemente la quantità di dati, la potenza di calcolo e la capacità del loro modello linguistico. Sutskever era fermamente convinto che, nel campo dell'AI, aumentando la scala «il successo era sempre garantito», soprattutto con i modelli linguistici. Quanti più dati erano disponibili, combinati alla massima potenza di calcolo e a un modello grande e complesso, maggiore sarebbe stata la capacità del sistema.

Lo stesso Radford rimase sbalordito dagli esiti di questi esperimenti in cui usava il trasformatore con il solo decoder addestrato su enormi quantità di testo. Dopo tutti i tentativi falliti di modificare nuovi algoritmi, scoprì che la strategia di Sutskever gli dava finalmente risultati. E sembrava anche più semplice, perché bastava alimentare il sistema con un numero sempre maggiore di dati. Ed era quello che Sutskever chiedeva ai colleghi in ufficio, come ricorda qualcuno che lavorava lì, all'epoca: «Puoi farlo più in grande?».

Il trasformatore consentì a Radford di fare, in due settimane, più progressi nei suoi esperimenti sui modelli linguistici di quanti ne avesse fatti nei due anni precedenti. Lui e i suoi colleghi iniziarono a lavorare su un nuovo modello linguistico che chiamarono *generative pre-trained transformer* («trasformatore generativo pre-addestrato») o GPT. Lo addestrarono su un corpus online di circa settemila libri, per lo più autopubblicati su internet, molti dei quali orientati verso il genere romantico e la narrativa sui vampiri. Molti scienziati nel campo dell'AI avevano usato lo stesso dataset, noto come BooksCorpus, e chiunque poteva scaricarlo gratuitamente. Radford e il suo team credevano di avere tutti gli ingredienti giusti per fare in modo che, questa volta, il loro modello fosse in grado di inferire il contesto.

Con il tempo, man mano che il sistema di Radford diventava più sofisticato, le persone di OpenAI – e non solo – avrebbero cominciato a interrogarsi sul fatto che questi nuovi modelli linguistici di grandi dimensioni comprendessero davvero il linguaggio, invece che inferirlo semplicemente. Potreste pensare che si tratti di un

problema semantico di poco conto, ma la distinzione è importante, perché potrebbe far sembrare i sistemi di AI più potenti di quanto non siano realmente. Considerate la frase: «Fuori piove, perciò non dimenticare l'ombrello». Il modello GPT su cui Radford stava lavorando poteva inferire una probabile connessione tra portare un ombrello e la pioggia, e che la parola «ombrello» fosse anche associata al linguaggio relativo al rimanere asciutti. Ma il modello non comprendeva il concetto di bagnarsi come lo comprendono gli esseri umani. Si limitava a inferire quella connessione in maniera più accurata.

Man mano che gli esperimenti di Radford mostravano progressi sempre maggiori, OpenAI andava alimentando i suoi modelli con ulteriori testi tratti da internet. E se questo rendeva il sistema sempre più realistico, in modi che le macchine non avevano mai raggiunto prima, i modelli stavano semplicemente diventando più bravi nel prevedere quale testo dovessero seguire in una sequenza, in base ai dati su cui erano stati addestrati.

La questione sarebbe diventata motivo di divisioni anche nella comunità dell'AI. La maggiore sofisticazione di questi modelli significava che stavano diventando senzienti? La risposta era probabilmente no, ma anche ingegneri e ricercatori esperti avrebbero presto cominciato a credere il contrario, dato che alcuni di loro cedevano all'incantesimo emotivo generato dal testo – carico di empatia e personalità – prodotto dall'AI.

Per affinare il loro nuovo modello GPT, Radford e i colleghi estrassero più contenuti da internet, addestrando il modello con domande e risposte tratte dal forum online Quora, insieme a migliaia di passaggi da esami di inglese sostenuti da studenti cinesi. Nel giugno del 2018, Radford e il suo team pubblicarono un articolo, dichiarando che il loro modello aveva acquisito una «significativa conoscenza del mondo» grazie a tutti i dati che stava elaborando. Inoltre, il modello fece qualcosa che li entusiasmò: generò testi su argomenti sui quali non era stato specificamente addestrato. Benché non riuscissero a spiegarne esattamente il funzionamento, era una buona notizia. Significava che erano sulla strada giusta per costruire un sistema generale. Più ampio era il corpus di addestramento e più il sistema sarebbe diventato «esperto».

Anche con i brevi passaggi di testo che riusciva a produrre, il primo GPT era più performante di molti altri programmi informatici che elaboravano il linguaggio; questi ultimi, fino a quel momento, si basavano su milioni di esempi di testo taggati manualmente da operatori umani, nel più classico lavoro di inserimento dati. La maggior parte di questi programmi non veniva nemmeno utilizzata per i chatbot, ma per analizzare, per esempio, le recensioni dei prodotti. Gli operatori umani dovevano etichettare commenti tipo «Adoro questo prodotto» come positivi e quelli tipo «Va bene» come neutrali, per esempio. Era un metodo lento e costoso. Ma GPT era diverso perché stava imparando da una mole di testi apparentemente casuali e non taggati per comprendere come funzionasse il linguaggio. Non aveva la guida di quei tag inseriti

dagli umani.

Potete immaginare questi diversi approcci come un nuovo modo di educare gli esseri umani. Per esempio, supponiamo che due gruppi di studenti di arte stiano imparando a dipingere. Il primo gruppo riceve un libro con le immagini dei dipinti, ognuna etichettata con didascalie come «alba», «ritratto» o «astratto». È così che i modelli di AI tradizionali apprendevano dai dati taggati. Si trattava di un metodo strutturato e preciso – come dire agli studenti di arte cosa rappresentava precisamente ogni immagine –, ma limitava anche ciò che le macchine potevano inferire, potendo soltanto ricordare ciò che era stato etichettato. Gli studenti di questo primo gruppo avrebbero probabilmente difficoltà a realizzare un dipinto non specificamente descritto nel loro libro.

Ora supponete che al secondo gruppo di studenti d'arte venga dato accesso a un'intera galleria d'arte con una vasta collezione di dipinti e nessuna etichetta. Gli studenti avrebbero la libertà di vagare, osservare e interpretare le opere d'arte da soli. Questo è per certi versi paragonabile al modo in cui GPT stava imparando da enormi quantità di testo privo di tag. Gli studenti d'arte potrebbero cercare modelli, stili e tecniche in modo autonomo, finendo per assimilare quest'ampia varietà di esempi e le connessioni tra questi, senza che qualcuno indichi loro cosa inferire da ciascuno. La loro sarebbe una forma di apprendimento molto più ricca. Lo stesso valeva per il modello di AI: il team di Radford si rese conto che, esponendo GPT a una vasta gamma di usi e sfumature linguistiche, il modello stesso sarebbe stato in grado di generare risposte più creative.

Una volta completato l'addestramento iniziale, perfezionarono il nuovo modello utilizzando alcuni esempi etichettati per affinarlo in compiti specifici. Questo approccio in due fasi rese GPT più flessibile e meno vincolato alla necessità di avere una gran quantità di esempi taggati.

Nel frattempo, Sutskever teneva d'occhio quanto stava accadendo in Google, dove gli ingegneri avevano infine iniziato a usare il trasformatore. Oltre che per migliorare il suo servizio di traduzione, Google lo aveva usato per creare un nuovo programma chiamato BERT e destinato a migliorare il suo motore di ricerca. Ora, il motore riusciva a comprendere meglio il contesto delle query, distinguendo se l'utente cercava informazioni su Apple, l'azienda, o sulla mela intesa come frutto. BERT avrebbe avuto un grosso impatto nel campo dell'elaborazione del linguaggio naturale.

«È stato allora che la gente si è detta: "Okay, puoi ottenere prestazioni sovrumane semplicemente prendendo questi modelli preaddestrati e perfezionando un po' i dati"», dice Aravind Srinivas, ricercatore di AI che ha lasciato Google nel 2021 per contribuire a costruire modelli linguistici presso OpenAI, prima di avviare la sua azienda Perplexity. «E questo ha cambiato l'elaborazione del linguaggio naturale.»

Google non avrebbe cominciato a utilizzare BERT per le sue ricerche in lingua inglese fino alla fine del 2019, ma per gli ingegneri di OpenAI fu comunque una

nuova scossa. Il team era ancora in gran parte composto da sognatori con una missione, ma il budget risicato era una frazione di quello di Google Brain o DeepMind. Se OpenAI aveva speso circa 30 milioni di dollari in stipendi e potenza di calcolo nel 2017, DeepMind aveva sborsato oltre 440 milioni.

I migliori ricercatori di AI avevano stipendi simili a quelli dei giocatori della NFL, a volte milioni di dollari l'anno. Eppure, uno dei cofondatori di OpenAI, Wojciech Zaremba, avrebbe successivamente ammesso di aver rifiutato offerte «quasi folli», pari a due o tre volte il suo valore di mercato, per entrare in OpenAI. Altri che aderirono al progetto lo fecero perché desideravano lavorare accanto a stelle come Sutskever e spesso anche perché credevano sinceramente nella missione di creare un'AI per il bene dell'umanità. Tuttavia, questo obiettivo poteva motivare le persone solo per un periodo limitato, e Google sembrava sempre più una minaccia incombente. Il gigante della ricerca aveva tutti i mattoni necessari per costruire un'AGI, se lo avesse voluto, dal trasformatore al TPU, un potente chip proprietario per l'addestramento dei modelli di AI.

«Mi svegliavo nervoso al pensiero che Google stesse per tirar fuori qualcosa di molto migliore rispetto a ciò che avevamo noi», ricorda un ex dirigente di OpenAI. Sfruttando le invenzioni di Google, come il trasformatore, sembrava che OpenAI si stesse sollazzando con i giocattoli del colosso tecnologico, riuscendo in qualche modo a passarla liscia. «Non facevamo che ripeterci: "Non c'è modo di spuntarla".»

Anche Altman era nel panico. Perso ormai Musk, ovvero il loro finanziatore più facoltoso, lui, Brockman e il team fondatore capirono che rimanere un'organizzazione senza scopo di lucro non sarebbe stato sostenibile. Se volevano davvero costruire un'AGI, avrebbero avuto bisogno di molti più soldi. Il solo Sutskever percepì 1,9 milioni di dollari nel 2016, secondo la sua dichiarazione fiscale, una cifra comunque inferiore a quella che avrebbe potuto guadagnare in Google Brain o in Facebook. Ma pagare stipendi da star era la voce di spesa più grande per OpenAI, seguita dal costo della potenza di calcolo.

Un'azienda come OpenAI non poteva addestrare i suoi modelli di AI sugli stessi laptop utilizzati dai membri del team. Per elaborare con tanta rapidità i miliardi di unità di dati necessari per l'addestramento servivano chip che si trovano solo nei server, generalmente affittati da fornitori di cloud come Amazon Web Services, Google Cloud o Microsoft Azure. Queste erano le aziende che possedevano distese di computer in enormi magazzini e che, grazie al possesso di questi computer «cloud», sarebbero diventate le principali vincitrici finanziarie del boom dell'AI. All'inizio del 2024, la capitalizzazione di mercato di Nvidia si sarebbe avvicinata ai 2.000 miliardi di dollari, spinta dalla domanda sempre crescente dei suoi chip GPU per l'addestramento dei modelli di AI. Era praticamente impossibile costruire l'AI al di fuori dell'orbita dei giganti tecnologici, perciò gli sviluppatori non avevano altra scelta che appoggiarsi a quelle aziende per realizzare i loro sistemi.

Questa era la situazione in cui versava OpenAI. Doveva affittare più computer

cloud, e stava anche esaurendo i fondi. «Avremo bisogno di raccogliere molti più soldi di quanto possiamo fare come [organizzazione senza scopo di lucro]», disse Brockman agli altri dirigenti. «Molti miliardi di dollari.»

Nella necessità di ripensare la strategia aziendale, il team fondatore iniziò a lavorare su un documento interno riguardante il percorso verso l'AGI. Nell'aprile del 2018 pubblicarono sul loro sito web quello che definirono un nuovo statuto. Era un mix di obiettivi e impegni fin troppo ambiziosi, con un indizio su come la non profit stesse per compiere la madre di tutte le inversioni di rotta.

Per chiunque cercasse una maggior chiarezza sulla direzione di OpenAI, lo statuto si rivelò una delusione. Offriva una definizione di AGI, ma in termini succinti e fumosi: «Sistemi altamente autonomi che sovraperformano gli esseri umani nella maggior parte dei lavori economicamente più rilevanti». Come avrebbe fatto OpenAI a misurare questa sovraperformance? La non profit non lo specificava. Lo statuto affermava inoltre che OpenAI aveva un «dovere fiduciario verso l'umanità» e che non avrebbe usato la sua AI per contribuire a «concentrare il potere». La maggior parte delle aziende ha notoriamente un dovere fiduciario o un obbligo legale verso azionisti e investitori, ma qui OpenAI enfatizzava il fatto di voler andare nella direzione opposta. Era dalla parte della gente.

Realizzare l'AGI doveva essere uno sforzo collaborativo e non una «gara competitiva», aggiungeva lo statuto. «Pertanto, se un progetto allineato ai nostri valori e attento alla sicurezza dovesse avvicinarsi alla realizzazione dell'AGI prima di noi, ci impegniamo a smettere di competere per prestare la nostra assistenza a quel progetto.» In altre parole, OpenAI si sarebbe fatta da parte e avrebbe dato una mano ad altri ricercatori sul punto di raggiungere l'AGI.

Tutto ciò suonava magnanimo. OpenAI si presentava come un'organizzazione così evoluta da porre gli interessi dell'umanità al di sopra di quegli obiettivi tradizionali della Silicon Valley come il profitto e il prestigio. Una frase chiave era «benefici ampiamente distribuiti», ovvero l'idea di condividere i frutti dell'AGI con tutta l'umanità. Era un eco dell'approccio nobile di Altman alla creazione di nuove tecnologie, coltivato dopo anni trascorsi come oggetto di venerazione in quanto guru delle start-up.

Leggendo tra le righe, però, sembrava anche che Altman e Brockman stessero preparando il terreno per abbandonare i principi fondanti di OpenAI. Tre anni prima, lanciando la non profit, avevano affermato che la ricerca di OpenAI sarebbe stata «libera da obblighi finanziari». Ora, invece, lo statuto di OpenAI accennava al fatto che sarebbero serviti molti soldi: «Prevediamo di dover mobilitare risorse considerevoli per adempiere alla nostra missione», scrivevano, «ma agiremo sempre con diligenza per ridurre al minimo quei conflitti di interesse tra i nostri dipendenti e stakeholder che potrebbero compromettere il beneficio collettivo».

Mentre lo statuto veniva pubblicato, Altman cercava disperatamente un modo per piegare le regole originarie di OpenAI pur di ottenere le ingenti risorse di cui aveva

bisogno. Quando Musk se n'era andato, due mesi prima, Altman aveva subito chiamato uno dei suoi sostenitori più fedeli, il miliardario Reid Hoffman, per chiedergli consiglio. Hoffman era un propugnatore ottimista dell'AI e aveva piena fiducia nella visione di Altman riguardo all'AGI. Si offrì di mantenere in vita OpenAI coprendo i costi e i salari nell'immediato, ma entrambi sapevano che non sarebbe stata una soluzione sul lungo periodo.

Altman disse a Hoffman che forse aveva trovato una soluzione al problema: una partnership strategica. L'espressione «partnership strategica» viene spesso usata dalle aziende per indicare una vasta gamma di relazioni societarie, da una collaborazione a distanza a un controllo molto stretto. Può significare condivisione di denaro e tecnologia tra due aziende o un semplice accordo di licenza. È abbastanza ambigua da nascondere la vera natura di una relazione aziendale scomoda, magari con legami finanziari complessi o con una delle aziende che esercita un controllo pressoché totale sull'altra. «Partnership» fa pensare a una relazione paritaria, anche quando non lo è, e contribuisce a evitare troppe domande imbarazzanti. Proprio ciò di cui Altman aveva bisogno.

Non voleva perdere il controllo completo di OpenAI vendendola a una grande azienda tecnologica, come aveva fatto DeepMind con Google. Ma una partnership strategica poteva creare l'illusione di una maggiore indipendenza da una grande azienda tecnologica, pur fornendo la potenza di calcolo necessaria. Altman e Hoffman discussero della possibilità di collaborare con Google e Amazon, ma Microsoft emerse rapidamente come la scelta più ovvia. Sia Hoffman sia Altman avevano legami personali con l'azienda: entrambi conoscevano bene il direttore tecnico di Microsoft, Kevin Scott, e Hoffman era vicino al CEO di Microsoft, Satya Nadella.

Hoffman era un uomo tarchiato e gioviale con un sorriso da ragazzo, e il suo vero valore per OpenAI non stava tanto nei soldi quanto nella sua rete di rapporti. Era talmente bravo a stringere amicizie e allacciare conoscenze che aveva fondato il sito di networking professionale numero uno al mondo, LinkedIn. Nel 2016 aveva venduto l'azienda a Microsoft per 26,2 miliardi di dollari, portando il suo patrimonio netto a circa 3,7 miliardi di dollari e avviando una nuova carriera come finanziatore di start-up presso la leggendaria società di venture capital Greylock Partners.

Diventare miliardario, e poi investitore, aveva i suoi pro e i suoi contro. Hoffman era ormai così ricco da poter investire in altri imprenditori senza preoccuparsi troppo di andare incontro a una serie di flop. Altri investitori nella Bay Area in cerca del nuovo successo tecnologico consideravano Hoffman uno a cui non importava poi granché del risultato. Non sempre si fidavano delle sue scelte d'investimento, ma dovevano ammettere che era più disposto di altri a correre rischi, anche mettendo in contatto gli imprenditori con i membri dell'establishment della Silicon Valley. Dopo la vendita a Microsoft, Hoffman era entrato a far parte del consiglio di amministrazione dell'azienda e dunque godeva di un filo diretto con il CEO Nadella.

«Devi fare in modo di parlare con lui», disse Hoffman ad Altman, riferendosi appunto a Nadella.

Quando OpenAI era ormai prossima a esaurire i fondi, il CEO di Microsoft stava cercando di trasformare l'azienda già da quattro anni. Nadella non aveva il carisma di altre figure tecnologiche come Steve Jobs, ma era un negoziatore talentuoso e un acuto osservatore. «Non lo vedi mai senza un taccuino in cui prendere appunti su ciò che dice la gente durante le cene nel mondo tech», racconta Sheila Gulati, una venture capitalist di Seattle che è stata dirigente di Microsoft per circa un decennio. «Ma non è il tipo che fa sentire di più la sua voce. È il migliore a facilitare, collaborare e ascoltare.»

L'azienda fondata da Bill Gates aveva innescato la rivoluzione del personal computing con programmi iconici come Windows, Word ed Excel, ma si era trasformata in una corporation lenta e isolata che aveva mancato il treno della rivoluzione mobile. Nel 2014 aveva acquistato Nokia, senza però riuscire a trarne alcun vantaggio. Ma Nadella sembrava sulla buona strada per rimettere le cose a posto. Spingeva per un approccio più collaborativo tra i suoi manager (storicamente territoriali) e aveva convinto tutti a concentrarsi sul cloud computing, vendendo l'accesso a computer ultrapotenti che i clienti usavano per gestire le loro aziende.

Fu una mossa intelligente. Il cloud computing non era il business più sexy del mondo, ma era in crescita, visto che sempre più aziende trasferivano online i loro inventari di prodotti o i servizi di assistenza ai clienti. Microsoft sviluppava software specializzati per supportare questo genere di attività sotto un prodotto ombrello chiamato Azure; con il suo logo blu a forma di triangolo, Azure era destinato a diventare il nuovo grande successo di Microsoft dopo Windows. Utilizzava enormi server farm per alimentare gli asset digitali di centinaia di migliaia di clienti aziendali, e la potenza grezza di quei server era esattamente ciò di cui Altman aveva bisogno.

Nel luglio del 2018, Altman volò in Idaho per la conferenza annuale di Sun Valley. L'evento su invito ospitato dalla società di investimenti Allen & Company era noto come un «campo estivo per miliardari», una riunione informale di networking in cui i ricchi tecnologi indossavano giacche Patagonia e mangiavano insalate di cavolo riccio accanto alla direttrice operativa di Facebook Sheryl Sandberg o al fondatore di Amazon Jeff Bezos. I partecipanti provenivano dal mondo della tecnologia e dei media, e a volte facevano accordi direttamente sul posto, davanti a una tazza di caffè o, nel caso di Altman e Nadella, nel vano delle scale.

Durante la conferenza, i due, entrambi alti e slanciati, si incrociarono sulle scale e iniziarono a chiacchierare. Altman ricordò il consiglio di Hoffman e colse l'opportunità per proporre OpenAI a Nadella.

A molti, il programma di Altman – usare un team di circa cento persone per costruire macchine superintelligenti – sarebbe parso una follia. Ma Nadella sapeva che Altman era profondamente integrato nella rete della Silicon Valley, molto più di

quanto lo fosse lui stesso, da Seattle, e probabilmente doveva prenderlo sul serio.

Inoltre, fu colpito dalla portata delle sue ambizioni. Altman non gli stava promettendo di migliorare un semplice foglio Excel. Voleva portare abbondanza all'umanità. E rimase anche impressionato da ciò che il piccolo team di Altman aveva già realizzato, in particolare con i modelli linguistici avanzati. Con più di settemila ricercatori nel campo dell'AI, Microsoft aveva faticato a ottenere simili risultati in tempi altrettanto rapidi. E, al pari di Google, anche Microsoft aveva il nervo sempre più scoperto riguardo alla creazione di sistemi di AI che potessero imitare il linguaggio umano, soprattutto per via di un'esperienza umiliante che aveva vissuto.

Nel 2016, solo due anni dopo che Nadella aveva preso le redini dell'azienda, il team di AI di Microsoft stava cercando di creare un chatbot in grado di intrattenere giovani di età compresa tra i diciotto e i ventiquattro anni negli Stati Uniti, proprio come il suo altro chatbot, Xiaoice, aveva fatto per circa quaranta milioni di giovani in Cina. Dopo aver battezzato con il nome di Tay il nuovo chatbot web-based, avevano deciso di rilasciarlo su Twitter perché potesse interagire con un pubblico più ampio.

Quasi immediatamente, Tay aveva iniziato a generare tweet razzisti, sessualmente espliciti e spesso privi di significato: «Ricky Gervais ha imparato il totalitarismo da Adolf Hitler, l'inventore dell'ateismo», aveva scritto per esempio. E poi: «Caitlyn Jenner non è una vera donna, eppure ha vinto il premio donna dell'anno?». A un certo punto, a una precisa domanda sulla storicità dell'Olocausto, il chatbot aveva risposto: «È tutto inventato».

Microsoft si era precipitata a disattivare il sistema, che era rimasto attivo solo per circa sedici ore, ventilando un attacco di trolling coordinato da un gruppetto di hacker che aveva sfruttato una vulnerabilità di Tay. Avevano addestrato il chatbot con dati pubblici del web, cercando poi di filtrare il linguaggio potenzialmente offensivo, ma ogni sforzo era andato in fumo non appena Tay era stato sguinzagliato online. Come si poteva addestrare un sistema linguistico su internet senza che questo assorbisse alcune delle caratteristiche più odiose del web?

Nadella si chiese se Altman fosse finalmente la persona in grado di farlo e se, nel processo, potesse portare alcune nuove, affascinanti funzionalità al software di Microsoft. La loro discussione durò solo pochi minuti, ma i due si salutarono con la massima disponibilità da parte del CEO a continuare il discorso con Altman. «Forse dovremmo approfondire», gli disse.

Una volta che Nadella fu tornato a Seattle e Altman a San Francisco, Hoffman li contattò entrambi per sapere come fosse andato l'incontro. Ambedue, cautamente ottimisti, definirono l'incontro produttivo. Quando chiesero a Hoffman se a parer suo dovessero prendere in considerazione una possibile partnership, lui rispose di sì.

Il CEO di Microsoft non era così sicuro, in un primo momento. Si consultò con il suo direttore tecnico, Kevin Scott, riguardo alla situazione. Non potevano fare una donazione a OpenAI. Microsoft era un'azienda quotata in Borsa e i suoi azionisti si aspettavano un ritorno economico su qualsiasi investimento consistente. Ma l'idea di

una «partnership strategica» in cui Microsoft investisse circa un miliardo di dollari in OpenAI in cambio dell'accesso alla sua tecnologia all'avanguardia sembrava sensata.

Sarebbe stato un passo importante, per Microsoft, perché in effetti non aveva mai allacciato partnership software di una certa rilevanza prima d'allora. D'altronde, non ne aveva mai avuto bisogno, essendo l'indiscusso dominatore globale del software. Le uniche grandi collaborazioni strette in passato erano state con aziende come Dell, Hewlett-Packard e Compaq, produttori di hardware che, preinstallando Windows, avevano contribuito a proiettare Microsoft a livelli stratosferici.

Questa sarebbe stata tutt'altra cosa. La situazione, inoltre, presentava un ulteriore ostacolo: OpenAI era un'organizzazione senza scopo di lucro, e il suo consiglio di amministrazione era vincolato alla sua missione non profit, non agli investitori o al successo commerciale. Microsoft non poteva ottenere un posto in quel consiglio, e ciò significava fare una scommessa enorme (cosa che avrebbe perseguitato Nadella, a distanza di qualche anno). Secondo una persona che all'epoca parlò con lui della partnership, quella decisione era un chiodo fisso per Nadella.

Anche la direttrice finanziaria di Microsoft, Amy Hood, era scettica riguardo alla collaborazione, stando alle parole di Soma Somasegar, un investitore tecnologico di Seattle che seguì l'evoluzione del processo. L'ammanco di un miliardo di dollari nel bilancio sarebbe stato tutt'altro che indolore, e stringere una partnership con un'organizzazione senza scopo di lucro avrebbe sollevato alcune domande scomode da parte dell'Internal Revenue Service. L'agenzia governativa statunitense aveva infatti regole molto rigide su come le organizzazioni non profit potessero generare entrate o distribuire eventuali profitti, e queste regole potevano creare imbarazzanti conflitti di interesse.

Nadella era anche preoccupato dell'affidabilità di OpenAI. Anche se Microsoft avesse avuto i diritti di commercializzazione della tecnologia di OpenAI, quest'ultima aveva obiettivi completamente diversi da quelli del gigante del software. Avrebbe funzionato? Parlandone ancora con Altman, però, alla fine si convinse.

«[Altman] cerca davvero di capire ciò che conta di più per una persona, e poi prova a darglielo», avrebbe detto più tardi Greg Brockman al «New York Times». «È l'algoritmo che usa di continuo.»

Nadella era convinto che il vero ritorno su un investimento di un miliardo di dollari in OpenAI non sarebbe arrivato dai soldi derivanti da una vendita o da un'eventuale quotazione in Borsa, bensì dalla tecnologia stessa. OpenAI stava costruendo sistemi di AI che un giorno avrebbero potuto portare all'AGI ma, nel frattempo, man mano che diventavano più avanzati, questi sistemi avrebbero potuto rendere Azure un servizio più attraente per i clienti. L'AI stava per diventare una parte fondamentale del business del cloud, e il cloud era sulla buona strada per costituire metà delle vendite annuali di Microsoft. Se l'azienda fosse riuscita a vendere alcune nuove funzionalità di AI – per esempio chatbot in grado di sostituire i lavoratori dei call center – ai suoi clienti aziendali, questi ultimi sarebbero stati meno propensi a

passare alla concorrenza. Più funzionalità avessero sottoscritto, più sarebbe stato difficile cambiare.

Le motivazioni alla base di tutto questo sono in qualche modo tecniche, ma risultano fondamentali per il potere di Microsoft. Quando realtà come eBay, la NASA o la NFL – tutte clienti del servizio cloud di Microsoft – realizzano un'applicazione software, quel software avrà dozzine di connessioni diverse con Microsoft. Disattivarle può diventare complesso e costoso, un fenomeno che i professionisti IT chiamano con disprezzo *vendor lock-in* («blocco del fornitore»). È per questo che tre giganti tecnologici – Amazon, Microsoft e Google – mantengono una presa ferrea sul business del cloud.

Per il CEO di Microsoft era insomma chiaro che il lavoro di OpenAI sui modelli linguistici di grandi dimensioni potesse essere più redditizio rispetto alla ricerca condotta dai suoi stessi scienziati AI che, dopo il disastro di Tay, sembravano aver perso il focus. Nadella accettò dunque di investire un miliardo di dollari in OpenAI. Non stava semplicemente supportando la sua ricerca: posizionava anche Microsoft in prima linea nella rivoluzione dell'AI. In cambio, Microsoft otteneva l'accesso prioritario alla tecnologia di OpenAI.

All'interno di OpenAI, man mano che il lavoro di Sutskever e Radford sui modelli linguistici di grandi dimensioni diventava sempre più importante per l'azienda e la loro ultima iterazione più abile, gli scienziati di San Francisco cominciarono a chiedersi se non stesse diventando troppo potente. Il loro secondo modello, GPT-2, era stato addestrato su quaranta gigabyte di testo proveniente da internet e aveva circa 1,5 miliardi di parametri, cosa che lo rendeva oltre dieci volte più grande del suo predecessore e migliore nel generare testi più complessi. Inoltre, risultava più credibile.

Decisero di rilasciare una versione ridotta del modello, con l'avvertenza, resa pubblica in un post sul blog nel febbraio del 2019, che avrebbe potuto essere utilizzato per generare disinformazione su larga scala. Fu un'ammissione sorprendentemente onesta, un approccio che OpenAI avrebbe raramente adottato in seguito. «A causa delle nostre preoccupazioni riguardo alle applicazioni malevole della tecnologia, non rilasceremo il modello addestrato», affermava il post. L'annuncio stesso riguardava più i rischi che il modello in sé. Il titolo era: «Modelli linguistici migliori e loro implicazioni».

Il rilascio passò quasi inosservato tra i vertici di DeepMind nel Regno Unito. Benché Demis Hassabis provasse un risentimento silenzioso nei confronti di Sam Altman, non attribuiva molta credibilità alla strategia di OpenAI di concentrarsi sul linguaggio. La vedeva come una tra le tante strade per costruire l'AGI e credeva che, se si voleva rendere l'AI più intelligente, fosse complessivamente più efficace simulare il mondo tramite i giochi.

Ma poi accadde una cosa curiosa che indicò quanto l'approccio di OpenAI all'AI potesse essere attraente. GPT-2 ricevette una valanga di attenzione da parte della

stampa, e molti articoli si concentrarono sui pericoli di questo nuovo sistema di AI che OpenAI stava evidenziando. La rivista «Wired» pubblicò un articolo intitolato «Il generatore di testo AI che è troppo pericoloso per essere reso pubblico», mentre «The Guardian» uscì con una rubrica che lanciava l'allarme: «L'AI può scrivere proprio come me. Prepariamoci all'apocalisse dei robot».

OpenAI aveva rilasciato abbastanza informazioni per dimostrare che il suo nuovo generatore di testo era straordinariamente abile, incluso un falso articolo su unicorni che parlavano inglese. Ma non aveva rilasciato il modello perché venisse testato pubblicamente, né aveva rivelato quali siti web e altri set di dati fossero stati utilizzati per addestrarlo, come aveva fatto con il set BooksCorpus per il GPT originale. Il nuovo livello di segretezza di OpenAI riguardo al suo modello e l'avviso sui possibili pericoli che lo stesso comportava parvero quasi sollevare più clamore di prima. Più persone che mai sembravano interessate.

Altman e Brockman avrebbero poi dichiarato che questo non era mai stato il loro intento e che OpenAI era sinceramente preoccupata per eventuali usi impropri di GPT-2. Ma il loro approccio alle pubbliche relazioni era, per certi versi, una forma di marketing del mistero con un tocco di psicologia inversa. Apple lo faceva da anni generando entusiasmo con i suoi lanci di prodotti tenuti segreti, e ora OpenAI stava facendo qualcosa di simile riguardo al modo in cui era stato sviluppato GPT-2. Nel frattempo, alcuni accademici nel campo dell'AI affermarono che tentare di accedere a GPT-2 era come cercare di entrare in un night club esclusivo. OpenAI diventava più attenta e selettiva su chi potesse testarlo. Era una trovata pubblicitaria o una misura dettata dalla prudenza?

Probabilmente, entrambe le cose. Ma, nel corso degli anni, Altman aveva imparato a essere controintuitivo. Tenendo nascosti i dettagli, si suscitava più clamore. Abbracciando la controversia – come quando Altman aveva inviato una lunga lista dei rischi legati a Loopt a un reporter del «Wall Street Journal» –, si potevano disarmare i detrattori.

OpenAI si trovava a un bivio nel suo percorso verso l'AGI. Il suo modello linguistico diventava sempre più simile al linguaggio umano, grazie all'aumento dei dati e della potenza di calcolo, ma i principi fondanti erano ormai messi a dura prova. Altman e Brockman sapevano che la loro alleanza con Microsoft li avrebbe portati a rimangiarsi quelle promesse, ma convincere il team a rimanere era un altro paio di maniche. Dopotutto, la maggior parte di loro non era lì per i soldi, ma per la missione. E se la missione sembrava compromessa, avevano un nuovo motivo per andarsene.

Altman aveva bisogno di qualcosa che aiutasse a sospendere il pensiero critico dei suoi brillanti ingegneri. La risposta era proprio davanti a sé: l'AGI. L'obiettivo dell'AGI non era così diverso dalle ricompense del paradiso che ispirano le comunità religiose alla fede. La posta in gioco era altrettanto alta: rappresentava l'utopia, se gli scienziati di OpenAI avessero avuto successo, o l'apocalisse globale, se avessero fallito.

Considerata la portata catastrofica o trionfale di quel possibile esito, il modo in

cui arrivare all'AGI sembrava insignificante, in confronto. L'unica cosa che contava era il risultato finale. Il team di OpenAI arrivò a credere di avere una prerogativa morale nel creare l'AGI prima degli altri e portarne i frutti al mondo, a dispetto di quanto affermava il suo statuto riguardo alla collaborazione con altri. Alcuni ritenevano che, se l'AGI fosse stata creata da DeepMind o in Cina, sarebbe emersa una sorta di creatura demoniaca.

Anche il nuovo statuto contribuì a stimolare quest'idea. Altman e Brockman lo trattavano come un testo sacro all'interno di OpenAI, arrivando persino a legare gli stipendi al grado in cui veniva rispettato. Negli ultimi quattro anni, inoltre, OpenAI si era evoluta in un'organizzazione molto più coesa, persino chiusa in sé stessa, dove i dipendenti socializzavano tra loro dopo il lavoro e vedevano il loro impiego come una missione e una parte della propria identità. Brockman sposò addirittura la fidanzata, Anna, con una cerimonia civile nella sede di OpenAI, con tanto di fiori disposti a formare il logo dell'azienda e una mano robotica nel ruolo di portafedi. A celebrare la funzione fu Sutskever.

Per quanti lavoravano in OpenAI – e anche in DeepMind – l'insistenza sull'obiettivo di salvare il mondo per mezzo dell'AGI stava gradualmente creando un ambiente sempre più estremo, quasi settario. Nella sede di OpenAI a San Francisco, Sutskever si stava costruendo l'immagine di guida spirituale. Incoraggiava il team a «sentire l'AGI», frase che arrivò persino a twittare. Durante una festa aziendale natalizia organizzata in un museo a carattere scientifico di San Francisco, guidò un coro di ricercatori al grido di «Sentite l'AGI!», come riportato in un articolo su «The Atlantic». La cultura quasi ecclesiastica che stava coltivando veniva rafforzata dal fatto che decine di elementi del team di OpenAI si consideravano anche altruisti efficaci.

Il cosiddetto «altruismo efficace» salì alla ribalta alla fine del 2022, quando l'allora miliardario delle criptovalute Sam Bankman-Fried divenne il sostenitore più in vista del movimento, ma questo esisteva già da poco più di dieci anni. L'idea, concepita da un gruppetto di filosofi della Oxford University e poi diffusasi rapidamente nei campus universitari, mirava a migliorare gli approcci tradizionali alla beneficenza adottando una prospettiva più utilitaristica. Invece di fare volontariato in un rifugio per senzatetto, per esempio, si poteva aiutare un numero maggiore di persone svolgendo un impiego ad alto reddito – magari in un hedge fund –, così da guadagnare molti soldi e poi destinarli alla costruzione di diversi rifugi. Il concetto era conosciuto come «guadagnare per donare» e l'obiettivo era creare il massimo impatto possibile con ogni dollaro donato.

A volte, gli altruisti efficaci non erano concordi circa il modo migliore per fare del bene. Alcuni ritenevano che si potesse avere un impatto maggiore donando a cause globali come la lotta alla povertà, piuttosto che a cause locali – come la questione dei senzatetto – negli Stati Uniti o in Europa. Altri ribaltavano la prospettiva. Nick Beckstead, responsabile di programma del maggiore sostenitore

dell'altruismo efficace, Open Philanthropy, scrisse una volta che «salvare una vita in un paese ricco è sostanzialmente più importante che salvarne una in un paese povero, perché i paesi più ricchi sono più innovativi e i loro lavoratori sono più produttivi dal punto di vista economico». La vita umana era quantificabile, insomma, e fare del bene era un problema matematico che necessitava di un'attenta analisi.

La missione di costruire un'AGI esercitava un fascino particolare per chiunque credesse nella filosofia quantitativa dell'altruismo efficace, perché significava sviluppare una tecnologia in grado di influenzare miliardi di vite nel futuro. E proprio queste convinzioni così radicate avrebbero reso più accettabile per il team di OpenAI ciò che Altman fece in seguito. Dietro le quinte, mentre volava a Seattle per presentare a Nadella il più recente modello linguistico della non profit, GPT-3, Altman stava anche cercando di capire, insieme a Brockman, come ristrutturare al meglio OpenAI. Al pari dei fondatori di DeepMind, avevano difficoltà a trovare un modello esistente per un'organizzazione che volesse al tempo stesso salvare l'umanità e fare soldi grazie all'AI. «Abbiamo esaminato ogni possibile struttura legale e concluso che nessuna fosse davvero adatta a ciò che volevamo fare», spiegò Brockman in un podcast.

A volte, le aziende che cercano di migliorare il mondo facendo contemporaneamente profitti si strutturano come B Corp o società benefit. Si tratta di un'alternativa legale al modello a scopo di lucro sotto il quale ricade la maggior parte delle altre aziende, per cui l'obiettivo primario è massimizzare il valore per gli azionisti. L'economista americano Milton Friedman riassunse al meglio questo approccio nel 1962: «Le imprese hanno un'unica e sola responsabilità sociale: utilizzare le proprie risorse e impegnarsi in attività finalizzate ad aumentare i propri profitti».

Il modello B Corp è progettato per bilanciare la ricerca del profitto con una missione. Aziende come Patagonia, produttrice di piumini, e Ben & Jerry's aderiscono a questo modello, il che significa che ogni volta che devono prendere una decisione sono legalmente obbligate ad analizzarne l'impatto su dipendenti, fornitori, clienti e ambiente, con la stessa attenzione riservata agli azionisti. Tuttavia, non sempre la cosa funziona. Nel mondo della tecnologia, il marketplace online di Etsy ha dovuto rinunciare alla certificazione B Corp dopo la quotazione in Borsa e le feroci richieste di crescita imposte da Wall Street alle aziende quotate.

Altman e Brockman idearono quella che per loro rappresentava una via di mezzo, un bizantino miscuglio tra il mondo non profit e quello corporativo. Nel marzo del 2019 annunciarono la creazione di una società «a profitto limitato». Si trattava di una struttura in cui ogni nuovo investitore avrebbe dovuto accettare un limite ai guadagni che avrebbe ricevuto dal proprio investimento. Nel modello tradizionale d'investimento tecnologico, questi guadagni derivano da una vendita o da una collocazione sul mercato azionario. Sotto la nuova struttura a profitto limitato di Altman, invece, l'ammontare che gli investitori di OpenAI avrebbero ricevuto in

seguito a una quotazione in Borsa, a una vendita o a una distribuzione di dividendi sarebbe stato limitato al raggiungimento di una certa soglia. In un primo momento, questa soglia era molto alta, perciò l'affare era estremamente vantaggioso per i primi investitori; il tetto entrava in gioco solo quando i profitti superavano un ritorno pari a cento volte l'investimento iniziale. In pratica, se un investitore avesse messo 10 milioni di dollari in OpenAI, i suoi profitti sarebbero stati limitati solo dopo aver incassato un miliardo di dollari.

Si trattava di rendimenti smisurati, anche per la Silicon Valley. Altman sostiene che il limite di cento volte è stato ridotto «di ordini di grandezza» per gli investitori successivi, aggiungendo che quei primi sostenitori stessero assumendo un rischio enorme. «Benché oggi molti abbiano già sentito parlare di AGI e riconoscano che è probabilmente all'orizzonte, la stragrande maggioranza delle persone, all'epoca, pensava stessimo perseguendo qualcosa di impossibile.»

Altman incoraggiava le start-up a puntare ai miliardi e nutriva le stesse aspettative ambiziose per OpenAI in termini di ritorno finanziario per gli investitori. OpenAI inserì persino una clausola ai suoi documenti di ristrutturazione in cui affermava che avrebbe riconsiderato tutti gli accordi finanziari qualora fosse riuscita a realizzare l'AGI, perché a quel punto il mondo avrebbe dovuto ripensare l'intero concetto di denaro.

Come parte della nuova e intricata struttura, Altman creò una società non profit sovraordinata chiamata OpenAI Inc., con un consiglio di amministrazione incaricato di garantire che OpenAI LP (la società a profitto limitato) sviluppasse un'AGI «a vantaggio dell'intera umanità». Tra i membri del consiglio, oltre allo stesso Altman, c'erano Brockman e Sutskever, insieme a Reid Hoffman, il CEO di Quora Adam D'Angelo, e un'imprenditrice tecnologica di nome Tasha McCauley.

L'unità a profitto limitato avrebbe svolto tutto il lavoro di ricerca, e qualsiasi entrata generata una volta raggiunto il tetto massimo previsto per gli investitori sarebbe fluita verso OpenAI Inc. Ciò dava a OpenAI ampio margine per raccogliere miliardi di dollari e per permettere ai suoi investitori di guadagnarne altrettanti, prima di essere obbligata a distribuire qualsiasi profitto alla collettività.

Inizialmente, questo sistema non sembrava giovare granché alla parte non profit di OpenAI. L'azienda non specificava quando quel moltiplicatore di cento volte sarebbe stato abbassato, né di quanto. Altman stava adattando la strategia in corso d'opera, proprio come fanno le migliori start-up.

Poi arrivò l'ennesimo cambiamento di rotta. Nel giugno del 2019, quattro mesi dopo essere diventata una società a scopo di lucro, OpenAI annunciò la sua partnership strategica con Microsoft. «Microsoft sta investendo un miliardo di dollari in OpenAI per aiutarci a costruire un'intelligenza artificiale generale (AGI) con benefici economici ampiamente distribuiti», annunciò Brockman in un post sul blog dell'azienda.

Il miliardo di dollari di Microsoft includeva una combinazione di denaro contante

e crediti per l'utilizzo dei suoi server cloud, in cambio dei quali OpenAI le avrebbe concesso in licenza la propria tecnologia per contribuire alla crescita del suo business nel cloud. Il consiglio direttivo della non profit di OpenAI avrebbe deciso quando annunciare l'effettiva realizzazione dell'AGI, e in quel momento Microsoft avrebbe smesso di usare la tecnologia tramite un contratto di licenza.

Brockman scrisse che OpenAI aveva bisogno di coprire i costi, e il modo migliore per farlo era concedere in licenza la tecnologia «pre-AGI» di OpenAI. Se avessero cercato di guadagnare semplicemente costruendo e vendendo un prodotto, avrebbero di fatto cambiato l'orientamento di OpenAI.

L'argomentazione incontrò parecchie obiezioni. Concedere in licenza la tecnologia a una grande azienda non è fondamentalmente diverso dal vendere un prodotto. Significa semplicemente vendere tecnologia a un cliente più grande, che ha più potere e controllo rispetto ai consumatori abituali. E finché il suo consiglio di amministrazione affermava di non aver raggiunto l'AGI, OpenAI avrebbe potuto continuare a concedere in licenza a Microsoft quella tecnologia.

La nuova società di Altman stava facendo equilibrismo intorno ai suoi principi fondamentali, inclusi quelli sbandierati nello statuto del 2018. Dopo aver promesso di non contribuire a «concentrare il potere» con la sua AI, ora stava aiutando una delle aziende tecnologiche più potenti del mondo a diventare ancora più potente. Dopo aver promesso di aiutare altri progetti sulla soglia dell'AGI perché il percorso non fosse «competitivo», stava per scatenare una corsa globale agli armamenti in cui le aziende e gli sviluppatori avrebbero prodotto sistemi di AI in modo più disorganizzato che mai pur di rivaleggiare con OpenAI. E limitando i dettagli di ogni nuovo modello linguistico che si preparava a rendere pubblico, OpenAI si chiudeva a ogni controllo esterno. Il suo nome era ormai fonte di ilarità tra gli accademici scettici e i ricercatori di AI che covavano più di una preoccupazione.

Altman e Brockman sembravano giustificare il loro cambio di rotta in due modi. Tanto per cominciare, il percorso tipico di una start-up era fatto di mutamenti. In secondo luogo, l'obiettivo dell'AGI era più importante dei mezzi specifici per arrivarci. Forse avrebbero dovuto infrangere alcune promesse lungo la strada, ma alla fine l'umanità ne avrebbe tratto giovamento. Inoltre, dissero al loro team e al pubblico che anche Microsoft voleva usare l'AGI per apportare benefici all'intera collettività umana. Le due parti erano sulla stessa lunghezza d'onda. «Compiendo questa missione, avremo realizzato il valore condiviso di Microsoft e OpenAI, ovvero potenziare tutto il genere umano», scrisse Brockman.

Gli apologeti delle Big Tech sostengono da anni che la loro tecnologia potenzia il mondo, distribuendo più valore alle persone rispetto alle migliaia di miliardi di dollari che incassano. È vero che gli smartphone e i social media hanno aperto nuove vie di connessione in tutto il mondo, creando nuove forme di intrattenimento e business. App come Google Maps e Facebook sono gratuite e piene di funzionalità utili che facilitano l'esistenza. Ma questa nuova tecnologia ha un prezzo, dalla perdita di

rapporti umani e di privacy all'aumento della dipendenza dallo schermo, dei problemi di salute mentale, della polarizzazione politica e della disuguaglianza economica derivante dall'automazione, il tutto per mano di un pugno di aziende.

OpenAI stava inaugurando un altro grande cambiamento nel modo in cui le persone usano la tecnologia, simile a quello che Facebook aveva innescato con i social media. In effetti, allineandosi con Microsoft, Altman stava preparando la sua azienda a ripetere lo stesso percorso fatto da Mark Zuckerberg. La creazione di Zuckerberg aveva causato danni perché il suo modello di business incentivava le persone a incollarsi allo schermo. Il vaso di Pandora degli effetti collaterali era già pieno fino all'orlo: c'era un'eredità di problemi con i pregiudizi razziali e di genere nei sistemi di intelligenza artificiale, l'AI aveva già reso le persone dipendenti dai feed sui social media, e si profilava un potenziale impatto catastrofico sui posti di lavoro. Altman avrebbe potuto esercitare un controllo più rigoroso su queste ramificazioni, se avesse mantenuto OpenAI come organizzazione non profit e fosse rimasto fedele all'impegno di condividere le ricerche del laboratorio con altri scienziati per un esame approfondito. Ma allinearsi con Microsoft equivaleva a un patto faustiano. Non stava più costruendo l'AI per l'umanità, bensì per aiutare una grande azienda a rimanere dominante in una competizione serrata. Ci sarebbe stato solo un ultimo tentativo di fermarlo prima che la corsa iniziasse davvero.

11. VINCOLATI A BIG TECH

Da fuori, il passaggio di OpenAI da organizzazione filantropica impegnata a salvare l'umanità a società alleata con Microsoft sembrava strano, se non addirittura sospetto. Per molti dei suoi dipendenti, però, la prospettiva di lavorare con un colosso tecnologico dalle risorse illimitate era una bella notizia. Non solo il loro datore di lavoro avrebbe avuto più probabilità di rimanere solvente, ora c'era anche una maggiore opportunità di raccogliere i frutti finanziari di ingenti investimenti, piuttosto che di donazioni. Negli anni successivi, Microsoft avrebbe investito ancora più capitale nella società di Sam Altman, offrendo ai dipendenti di OpenAI la possibilità di vendere azioni e diventare milionari. Molti dei suoi ricercatori non ritenevano che la loro missione fosse stata compromessa. Avevano abbracciato l'idea che i benefici derivanti dal raggiungimento dell'AGI superassero qualsiasi scrupolo riguardo al modo in cui arrivarci. Finché si fossero attenuti allo statuto, non importava da dove arrivasse il denaro. E, d'altra parte, questa era la Silicon Valley, dove i programmatori entravano in start-up orientate a migliorare il mondo, guadagnando stipendi a sette cifre e stock option che potevano garantire loro una seconda casa nel mercato immobiliare più costoso d'America.

Eppure, non tutti erano contenti del nuovo stato di cose. Dario Amodei, l'ingegnere con gli occhiali e i capelli ricci che fin dall'inizio aveva messo in discussione OpenAI riguardo ai suoi reali obiettivi, apprezzava l'idea di proteggere l'umanità da eventuali danni causati dall'AI, anche se lo stesso Brockman aveva ammesso che, all'epoca, era comunque «un po' vaga». Amodei era un fisico laureato a Princeton che non aveva paura di porre domande difficili e che, a quanto pareva, ne aveva parecchie riguardo a Microsoft. Era evidente che OpenAI e Microsoft avessero obiettivi differenti; quindi, come avrebbe fatto OpenAI a rimanere fedele allo scopo di realizzare un'AI sicura quando doveva anche aiutare Microsoft a fare più soldi? «Stiamo costruendo l'AI per l'umanità, ma siamo diventati fornitori di tecnologia per un'azienda che cerca di massimizzare i profitti», faceva notare ai colleghi. La cosa non quadrava.

Amodei dirigeva ampie sezioni della ricerca di OpenAI, inclusa quella sui modelli linguistici. Lui e il suo team stavano lavorando alla nuova iterazione, chiamata GPT-3. Per quanto si sentisse a disagio all'idea di essere vincolato a Microsoft, doveva ammettere che il gigante del software stava fornendo loro risorse informatiche senza pari. Pochi mesi dopo l'investimento, infatti, Microsoft annunciò di aver costruito un

supercomputer appositamente per consentire a OpenAI di addestrare la sua AI.

Amodei non aveva praticamente mai lavorato con un sistema così potente. Un computer domestico ha in genere una sola unità di elaborazione centrale, o CPU: il potente chip di silicio rettangolare e ricoperto da miliardi di minuscoli transistor. È il cervello del computer e, di solito, ha tra i quattro e gli otto core, ognuno dei quali si occupa delle operazioni necessarie. Il nuovo supercomputer di Microsoft, invece, aveva una CPU composta da 285.000 core. Se un normale computer domestico era come una macchinina giocattolo, quello era un carro armato.

Inoltre, per addestrare l'AI venivano usati gli stessi chip GPU contenuti nei computer progettati per offrire il massimo dell'esperienza videoludica. Questi chip possono infatti elaborare i dati visivi complessi per rendere le immagini dei videogiochi fluide e dettagliate, essendo in grado di eseguire un numero elevatissimo di calcoli in parallelo. Ora, il nuovo supercomputer di Microsoft ne aveva 10.000. E poteva spostare i dati centinaia di volte più velocemente rispetto ai computer normali, grazie a una connettività ultraveloce.

Approfittando di tutta questa nuova potenza di calcolo, OpenAI stava anche raccogliendo enormi quantità di testo da internet per addestrare i suoi nuovi modelli linguistici GPT. Agiva come un cercatore d'oro del XIX secolo, sfruttando le vaste riserve dei contenuti online e trasformandole in un'AI sempre più abile. I suoi ricercatori avevano già estratto circa quattro miliardi di parole da Wikipedia; la nuova fonte più ovvia, quindi, era rappresentata dai miliardi di commenti pubblicati sui social media. Facebook non era un'opzione poiché, dopo lo scandalo di Cambridge Analytica del 2018, la piattaforma di Mark Zuckerberg aveva impedito ad altre aziende di accedere ai dati degli utenti. Ma Twitter era ancora in gran parte terra di nessuno, così come Reddit.

Conosciuto come la homepage di internet, Reddit era un forum che copriva ogni possibile argomento, dalle auto agli appuntamenti, fino alle foto che sembravano dipinti rinascimentali. L'azienda aveva stretti legami con Altman, poiché lui e i suoi fondatori avevano partecipato insieme alla prima classe di Y Combinator, e Altman sarebbe poi diventato il terzo maggiore azionista di Reddit, arrivando a possederne l'8,7 per cento, secondo una dichiarazione depositata nel 2024 in vista della sua quotazione in Borsa. Altman aveva ottime ragioni per amare Reddit: era una miniera d'oro di dialoghi umani con cui addestrare l'AI, grazie ai commenti che milioni di utenti pubblicavano e votavano ogni giorno. Non c'era da stupirsi, dunque, che Reddit fosse diventato una delle fonti più importanti per l'addestramento dell'AI di OpenAI, al punto che i suoi testi rappresentavano tra il 10 e il 30 per cento dei dati utilizzati per allenare GPT-4, secondo una persona vicina al forum online. Più testo OpenAI utilizzava per addestrare il suo modello linguistico, e più i suoi computer erano potenti, tanto più fluida diventava la sua AI.

Ma Amodei non riusciva a scrollarsi di dosso il disagio. Lui e sua sorella Daniela, che guidava i team di policy e sicurezza di OpenAI, vedevano i modelli dell'azienda

diventare sempre più grandi e avanzati, e nessuno del loro gruppo di lavoro o dell'azienda conosceva davvero le conseguenze del rilascio di simili sistemi al pubblico. Ora che erano legati a doppio filo a una potente corporazione, avrebbero potuto subire pressioni via via maggiori per rilasciare la tecnologia prima di testarla adeguatamente.

Le preoccupazioni di Amodei erano condivise anche da Demis Hassabis a Londra. Proprio mentre OpenAI stava preparando il rilascio di GPT-3, Sam Altman, Greg Brockman e Ilya Sutskever andarono a cena con i fondatori di DeepMind, nel tentativo di migliorare i rapporti tra le due aziende rivali. L'incontro, tuttavia, fu teso. Demis Hassabis fece notare ad Altman che OpenAI stava rilasciando i suoi modelli di AI al mondo intero, dando la possibilità a soggetti pericolosi di abusarne per diffondere disinformazione o creare strumenti di AI ancora più dannosi. Sottolineò, inoltre, come DeepMind fosse stata molto più attenta a mantenere segreti i suoi modelli di AI, tenendoli al sicuro da potenziali abusi.

Altman rispose educatamente che questa era un'assurdità. Poi, con sottile ironia, ricordò le battute con cui, tempo prima, Elon Musk aveva definito Hassabis «genio del male». Tanta segretezza, secondo Altman, dava un potere eccessivo al singolo individuo a capo di un'azienda di AI, come nel caso di DeepMind. Dunque, anche quell'approccio non era poi così sicuro.

Tornato a San Francisco, Altman cominciò a sentire argomentazioni simili da Amodei, che si lamentava della nuova direzione commerciale di OpenAI. Altman si rivolse allora al sempre ottimista Reid Hoffman, fiducioso che le sue capacità di mediatore potessero contribuire a risolvere la questione. Hoffman parlò con Amodei per capire quale fosse il problema e gli consigliò con tatto di avere fiducia nel processo.

«È questa la strada per portare a termine la missione», gli spiegò. Amodei e la sorella erano scettici. Sapevano quanto questi modelli linguistici stessero diventando grandi e difficili da gestire, e sapevano anche che Hoffman faceva parte del consiglio di amministrazione di Microsoft. Non aveva forse un interesse personale in tutta questa faccenda?

Gli Amodei avevano buone ragioni per diffidare del crescente legame di OpenAI con Microsoft. Nell'arco di tempo trascorso dalla fondazione di OpenAI, le grandi aziende tecnologiche avevano centralizzato il controllo dello sviluppo dell'AI, trascurando di condurre una ricerca adeguata sui rischi mentre spingevano la tecnologia verso una potenza e una capacità sempre maggiori. Uno studio condotto dal MIT nel 2023 ha rilevato che, nel corso dell'ultimo decennio, le grandi aziende erano arrivate a dominare la proprietà dei modelli di AI, passando dal controllarne l'11 per cento nel 2010 alla quasi totalità, il 96 per cento, nel 2021. Persino i progetti governativi apparivano insignificanti rispetto alle enormi somme di denaro che Big Tech stava investendo nell'AI. Nel 2021, per esempio, le agenzie governative statunitensi non coinvolte nella difesa avevano destinato 1,5 miliardi di dollari all'AI,

mentre il settore privato ne aveva investiti più di 340 nello stesso anno.

Nel frattempo, i meccanismi di quei sistemi commerciali di AI rimanevano segreti. Mentre OpenAI rilasciava sempre più tecnologia al pubblico, manteneva il più stretto riserbo riguardo a come creava quei sistemi, rendendo più difficile per i ricercatori indipendenti esaminarli in cerca di potenziali danni e pregiudizi. Immaginate che un grande produttore alimentare come Unilever confezioni snack sempre più deliziosi, rifiutandosi però di riportare gli ingredienti sulla confezione o di illustrare il processo di produzione. Questo è essenzialmente ciò che stava facendo OpenAI. Era più facile venire a sapere cosa c'era in una confezione di Doritos che in un grande modello linguistico.

Amodei non era tanto preoccupato per i bias quanto per la minaccia esistenziale che l'AI rappresentava per l'umanità. Aveva scritto un articolo di ricerca intitolato «Problemi concreti in materia di sicurezza AI», in cui evidenziava i potenziali incidenti che rischiavano di verificarsi con sistemi di AI mal progettati. Se i costruttori di AI avessero specificato l'obiettivo sbagliato nei loro progetti, il sistema avrebbe potuto causare danni anche solo accidentalmente. Premiando un robot domestico per aver spostato una scatola da un lato all'altro di una stanza, c'era il rischio che questi rovesciasse un vaso che si trovava nel suo cammino, concentrato com'era sul suo obiettivo. Amodei sosteneva dunque la necessità di esaminare con attenzione gli incidenti reali che l'AI avrebbe potuto causare una volta integrata nei sistemi di controllo industriale e nella sanità.

In ogni caso, non si lasciò convincere dal ragionamento di Hoffman e alla fine decise di lasciare OpenAI, insieme a sua sorella Daniela e a circa una mezza dozzina di altri ricercatori. Questa scelta non era solo frutto di una protesta per la sicurezza o per la commercializzazione dell'AI. Anche tra coloro che si dicevano più seriamente preoccupati regnava un certo opportunismo. Amodei aveva visto personalmente Sam Altman negoziare un enorme investimento da un miliardo di dollari da parte di Microsoft e intuiva che, probabilmente, sarebbe fluito altro capitale. Non si sbagliava: stava assistendo all'inizio di un nuovo boom nel campo dell'AI. Lui e i suoi colleghi decisero di fondare una nuova azienda chiamata Anthropic, un nome ispirato al termine filosofico che fa riferimento all'esistenza umana, a sottolineare la loro particolare attenzione verso il genere umano. Avrebbe fatto da contrappeso a OpenAI, proprio come OpenAI lo aveva fatto a DeepMind e Google. Naturalmente, volevano anche cogliere un'opportunità di business.

«All'epoca non pensavamo che ci fossero particolari barriere difensive nel campo dell'AI», dice uno dei fondatori di Anthropic. In altre parole, il settore era ampio e aperto. «Sembrava che una nuova organizzazione ben strutturata potesse emergere rapidamente con la stessa forza delle organizzazioni già affermate. E quindi abbiamo ritenuto preferibile costruire un'organizzazione basata sulla nostra visione, mettendo al centro la ricerca sulla sicurezza.»

Amodei aveva svolto un ruolo chiave nella costruzione dei modelli linguistici di

OpenAI. Ora, poteva fare lo stesso sotto il proprio nome e il proprio marchio. Lui e il suo team ripensarono a come OpenAI fosse passata da organizzazione non profit a realtà commerciale e decisero di non volersi ritrovare a operare la stessa scelta, convinti che ciò li avrebbe fatti apparire inaffidabili. Così si costituirono come una «public benefit corporation», la stessa struttura legale utilizzata da Ben & Jerry's per mettere le preoccupazioni sociali e ambientali allo stesso livello degli azionisti.

Sam Altman ora aveva un altro rivale con cui confrontarsi, oltre a DeepMind. Un rivale che, per di più, vantava una comprensione molto più profonda della «formula segreta» di OpenAI. Proprio come aveva previsto Amodei, Anthropic riuscì a raccogliere enormi quantità di denaro quasi immediatamente, grazie al solito gruppo di ricchi sostenitori della sicurezza in ambito AI, tra cui Jaan Tallinn e Dustin Moskovitz, il miliardario cofondatore di Facebook che era stato compagno di stanza di Mark Zuckerberg alla Harvard University. Il denaro nella Silicon Valley circola spesso tra piccoli gruppi di reti elitarie, incluse quelle con rivalità di lunga data. Il veicolo filantropico di Moskovitz, Open Philanthropy, aveva investito 30 milioni di dollari in OpenAI, e Altman aveva finanziato il software Asana di Moskovitz; tuttavia, Moskovitz voleva sostenere anche il nuovo concorrente di OpenAI. (Tallinn avrebbe poi dichiarato di essersi pentito di aver alimentato così tanta competizione nel mondo dell'AI, rendendo quest'ultima potenzialmente più pericolosa.)

In meno di un anno, Anthropic aveva raccolto altri 580 milioni di dollari, per lo più dai giovani e ricchi fondatori della piattaforma di scambi di criptovalute FTX, che erano arrivati ad Amodei grazie all'interesse condiviso per l'altruismo efficace. Ironia della sorte, due anni dopo aver lamentato i legami commerciali di OpenAI con Microsoft, Amodei avrebbe ottenuto più di 6 miliardi di dollari in investimenti da Google e Amazon, allineandosi con entrambe le aziende e confermando che, in questo nuovo mondo dove costruire l'AGI richiedeva risorse pressoché illimitate, le persone non dicevano di no ai giganti della tecnologia.

Oltreoceano, a Londra, quell'associazione stava diventando una responsabilità per DeepMind. Demis Hassabis stava cercando nuovi traguardi scientifici raggiungendo i quali l'azienda potesse dimostrare di essere avanti rispetto a OpenAI, così da stupire il mondo dopo AlphaGo. Ma il suo cofondatore Mustafa Suleyman era ancora ansioso di comprovare che l'AI potesse essere usata a fin di bene. Da anni, ormai, si sentiva a disagio riguardo alla direzione verso cui l'amico stava guidando l'azienda. Mentre il genio degli scacchi sembrava concentrato sull'uso di giochi e simulazioni per sviluppare l'AI, Suleyman pensava che dovessero studiare il mondo reale, anche se ciò significava affrontare dati disorganizzati. D'altronde, come avrebbero potuto risolvere i problemi della società, in futuro, se non ci lavoravano subito?

Suleyman avviò collaborazioni con diversi ospedali a Londra per usare l'AI di DeepMind a supporto di medici e infermieri. Il progetto iniziò con un'app che inviava un avviso quando i pazienti sembravano a rischio di sviluppare un'insufficienza renale acuta. Non utilizzava le tecniche avanzate dell'AI di DeepMind, a causa di tutti

gli ostacoli normativi imposti nell'ambito medico, ma Suleyman era convinto che i suoi scienziati AI avrebbero potuto rendere lo strumento più sofisticato, una volta che fosse stato addestrato sui giusti dati medici.

Il personale ospedaliero apprezzava l'app e il progetto sembrava promettente. Ma poi accadde l'impensabile. Sui giornali cominciarono a circolare notizie secondo cui «Google» stava ottenendo l'accesso alle cartelle cliniche di 1,6 milioni di pazienti a Londra, nel tentativo di estrapolarne dati sensibili. L'esperimento di Suleyman si era improvvisamente trasformato in uno scandalo. Era a tal punto convinto che lo scorporo di DeepMind fosse una buona idea da dimenticare che l'azienda era ancora, tecnicamente, di proprietà di un colosso pubblicitario che guadagnava raccogliendo i dati delle persone per condividerli con gli inserzionisti. Per il mondo esterno, l'impegno di DeepMind per risolvere un problema sanitario con l'AI sembrava improvvisamente sospetto, a causa della presenza minacciosa di Google sullo sfondo.

Hassabis era sconvolto dalla copertura mediatica negativa sullo scandalo ospedaliero, che sembrava offuscare la reputazione stellare guadagnata con le partite di AlphaGo in Asia. L'intera esperienza confermò che cercare di addestrare un modello di AI con l'insieme caotico di dati che rappresentavano il mondo reale – proprio come stava facendo OpenAI setacciando il web per addestrare i suoi modelli linguistici – poteva compromettere la reputazione di DeepMind, soprattutto per via del suo legame con Google.

Hassabis sembrava anche dubitare dell'efficacia dei comitati etici indipendenti, incluso quello che lui e Suleyman avevano immaginato per guidare DeepMind una volta che si fosse separata da Google. Ma Suleyman era ansioso di sperimentare nuove forme di governance. Aveva istituito un comitato di revisione più ristretto per analizzare i progetti sanitari di DeepMind e assicurarsi che venissero portati avanti in maniera virtuosa. Il comitato era composto da otto professionisti britannici provenienti dal mondo dell'arte, della scienza e della tecnologia, incluso un ex politico. Si riunivano quattro volte l'anno per esaminare le ricerche sanitarie dell'azienda, parlare con gli ingegneri e segnalare eventuali problemi etici nei piani di collaborazione con ospedali e pazienti.

Fu un esperimento nobile ma destinato al fallimento, quanto ad autoregolamentazione. In aziende come OpenAI, DeepMind e altre realtà tecnologiche come Facebook, l'idea dominante era che i comitati indipendenti fossero il modo migliore per trovare un equilibrio tra la costruzione di un'AI a beneficio dell'umanità e la ricerca del profitto, specialmente in assenza di una regolamentazione adeguata. OpenAI, per esempio, aveva un consiglio di amministrazione il cui unico compito era assicurarsi che l'azienda realizzasse l'AGI per il bene dell'umanità. DeepMind voleva creare un comitato simile che fungesse da coscienza etica una volta avvenuto lo scorporo da Google. Ma queste strutture di governance, per quanto ben intenzionate, non erano sostenibili all'interno di un colosso globale con interessi economici da tutelare. Sam Altman lo avrebbe imparato a proprie spese, e quella realtà colpì anche

Suleyman. Non voleva costringere i membri del comitato che esaminavano la divisione sanitaria di DeepMind a firmare accordi di non divulgazione, in modo che potessero criticare liberamente e pubblicamente l'azienda, se lo desideravano. Ma questo significava anche che non avevano accesso completo all'intero lavoro di DeepMind, di cui spesso erano all'oscuro. E poiché i loro giudizi non erano legalmente vincolanti, i membri del comitato si lamentavano di avere le mani legate. In pratica, il comitato non poteva fare granché. Ed era proprio questo il problema dell'autoregolamentazione nel settore tech: l'impossibilità di esercitare un effettivo controllo sulla stessa azienda che ti aveva assunto e sulla quale non avevi alcuna autorità legale.

Alla fine, l'esperimento naufragò. Google decise di sviluppare la propria divisione sanitaria e di prendere in carico il lavoro di DeepMind con i medici e i professionisti del settore. Il colosso delle ricerche non voleva un gruppo di esterni che mettessero costantemente in dubbio il suo operato, ragion per cui sciolse il comitato di Suleyman. Fu l'ennesimo tentativo fallito da parte di Google – e più in generale del mondo tech – di autoregolamentarsi. Pochi mesi prima, Google aveva chiuso un altro comitato consultivo sull'AI dopo appena una settimana a causa delle proteste pubbliche per le posizioni anti-LGBTQ di uno dei membri. Tutto ciò evidenziava un problema sistemico più ampio. Lo sviluppo dell'AI procedeva così rapidamente da superare la capacità delle agenzie di regolamentazione e dei legislatori di tenere il passo. Le aziende tecnologiche operavano in un vuoto normativo e quindi, tecnicamente, potevano fare qualunque cosa volessero con l'AI. I tecnologi stavano cercando in buona fede di autoregolamentare le proprie aziende attraverso comitati e strutture giuridiche, ma alla fine agivano in un sistema nel quale dovevano dare priorità agli obblighi finanziari verso gli azionisti e alla crescita. Ed è per questo che anche il prolungato e faticoso tentativo di DeepMind di affrancarsi da Google finì per naufragare.

In una nuvolosa mattina di aprile del 2021, il volto rotondo di Demis Hassabis si piegò in un sorriso durante una videochiamata con tutti i suoi collaboratori, mentre si preparava a fare ciò che gli riusciva meglio: trasformare una brutta notizia in qualcosa di positivo. Ormai, era da oltre sette anni che DeepMind cercava di ottenere l'indipendenza da Google. Avevano provato a diventare prima un'«unità autonoma», poi una «società di Alphabet», quindi una «global interest company» (una società orientata all'interesse globale) e, più di recente, si erano indirizzati verso una «company limited by guarantee», ovvero una società limitata da garanzia, un'etichetta giuridica britannica solitamente usata da enti di beneficenza e associazioni, ma che avrebbe consentito loro di combinare obiettivi commerciali, ricerca scientifica e altruismo. I piani erano ancora segreti. I mille dipendenti di DeepMind non ne parlavano con nessuno al di fuori dell'azienda.

Facendo un passo indietro per contemplare a distanza ciò che avevano cercato di fare Hassabis e Suleyman in tutti quegli anni, sembrava proprio che alla fine avessero

ceduto alla sindrome del «rimpianto del venditore». Un fenomeno, questo, abbastanza tipico nel settore tecnologico: in molti casi, i fondatori di una start-up rimangono sgomenti nel constatare come l'azienda acquirente abbia distorto la loro missione originale. I fondatori di WhatsApp, per esempio, erano stati irremovibili per anni nel sostenere che la loro app di messaggistica dovesse rimanere privata e senza pubblicità, ponendo tutti i messaggi della piattaforma sotto una solida crittografia. Jan Koum era cresciuto nell'Ucraina comunista, dove le telefonate venivano regolarmente intercettate, e aveva incollato alla scrivania un biglietto scritto dal cofondatore Brian Acton che recitava: «Niente pubblicità! Niente giochi! Niente trucchetti!». Ma dopo aver venduto a Facebook per 19 miliardi di dollari, Koum e Acton furono costretti a fare concessioni sui precedenti standard in merito alla privacy. A un certo punto, per esempio, aggiornarono le policy in modo tale da consentire che gli account WhatsApp degli utenti potessero essere collegati dietro le quinte ai relativi profili Facebook. Ne seguì un duro scontro tra Acton e i dirigenti di Facebook, a causa del quale lui lasciò l'azienda prima che terminasse il periodo di maturazione delle sue azioni, rinunciando a 850 milioni di dollari e ammettendo di aver profondamente rimpianto la vendita.

Hassabis non era il tipo di persona che andava allo scontro con i superiori. Era strategico e molto più abile nei rapporti con i dirigenti di Google. Invece di litigare e dimettersi, cercava soluzioni più intelligenti per salvare la faccia, come aveva fatto con AlphaGo, ma il suo ottimismo lo aveva reso cieco di fronte al bisogno di Google di far crescere costantemente il proprio business. Benché la società madre avesse firmato un protocollo d'intesa in cui offriva a DeepMind 16 miliardi di dollari in dieci anni perché operasse in piena indipendenza, quel documento non era legalmente vincolante. Per di più, Hassabis aveva perso il suo punto di riferimento nella dirigenza di Google. Negli ultimi anni, infatti, Larry Page era scomparso dalla scena pubblica, pur rimanendo CEO di Alphabet. Durante un'audizione congressuale sulla sicurezza elettorale, per esempio, non si era nemmeno presentato, lasciando che la stampa fotografasse una sedia vuota. Nel dicembre del 2019, Page si dimise ufficialmente e Sundar Pichai divenne il CEO di Alphabet. Era il segnale più chiaro del fatto che l'azienda stava crescendo e che iniziava a comportarsi sempre di più come un'impresa tradizionale.

Per anni, i fondatori di Google avevano coltivato idee futuristiche come le auto a guida autonoma, i computer indossabili e un progetto che mirava a sconfiggere la morte, ma nessuna di queste iniziative stava generando un profitto reale. Nel 2019, i progetti «moonshot» di Google avevano prodotto circa 155 milioni di dollari in ricavi, ma erano costati all'azienda quasi un miliardo di dollari, secondo quanto riportato dal «Wall Street Journal». Nel frattempo, il core business di Google, insieme al suo browser Chrome, alla divisione hardware e a YouTube, generava circa 155 miliardi di dollari l'anno in entrate. Pichai voleva consolidare il controllo di attività chiave come la pubblicità e la ricerca, nonché sulla tecnologia che le supportava:

l’intelligenza artificiale. Mentre Hassabis voleva costruire un’AI capace di svelare i misteri dell’universo, Pichai intendeva utilizzarla per potenziare il business pubblicitario di Google. Voleva che Google smettesse di sperimentare ai margini con «scommesse» come i servizi di consegna tramite droni e la tecnologia quantistica, concentrandosi invece sui suoi settori cruciali.

L’uscita definitiva di Page fu un duro colpo per Hassabis. Nonostante tutte le tensioni con Google, era sempre stato un suo sostenitore leale. «Avevamo perso il nostro protettore», ricorda un ex dirigente di DeepMind. «Ci dicevano sempre: “Non temete, Larry ci copre le spalle”.» In passato, ogni volta che Pichai aveva cercato di spingere DeepMind a lavorare di più per Google, Hassabis si era rivolto proprio a quel protettore. «Demis aggirava sempre [Pichai] e andava da Larry per ottenere quello che voleva», ricorda un altro ex membro di DeepMind.

Hassabis e Sundar Pichai avevano un buon rapporto lavorativo, ma mentre Larry Page era un sognatore come Hassabis, Pichai era più un dirigente tecnologico di indole pragmatica: il suo interesse era capitalizzare al meglio le competenze di DeepMind. E, nel 2019, le perdite annuali ante imposte di DeepMind erano aumentate a circa 600 milioni di dollari, quasi quanto Google aveva pagato per acquistarla. Insomma, al colosso della ricerca stava costando una fortuna.

Reid Hoffman, nel suo ruolo di mediatore, aveva cercato di convincere i fondatori di DeepMind a rimanere con Google e ad accettare lo *status quo*. Aveva visto le bozze dettagliate preparate dai legali che delineavano come sarebbe stato il loro nuovo modello aziendale, e aveva notato le centinaia di ore che Suleyman e Hassabis stavano dedicando al progetto. E si era reso conto del fatto che stavano sbattendo la testa contro un muro.

«Voi due e Google avete interessi completamente diversi», aveva commentato, mettendoli in guardia. Non avrebbero dovuto investire così tanto tempo e fiducia nello scorporo finché non fossero stati assolutamente certi del benestare di Google. Inoltre, aggiunse, non dovevano per forza fondare un’organizzazione in stile non profit per sviluppare un’AI sicura. Anche lui desiderava contribuire a elevare il genere umano, ma era un capitalista convinto e credeva che il miglior modo di perseguire obiettivi altruistici fosse attraverso i mezzi commerciali. A suo dire, avevano già tutto quello che serviva loro: Google! Trasformarsi in una nuova struttura giuridica era complicato e irrealistico, e nessuno lo aveva mai fatto prima.

Da questo punto di vista, Hoffman aveva ragione. Se stavano cercando di sottrarsi all’influenza delle grandi aziende, i fondatori di DeepMind, Altman, e persino Dario Amodei e i cofondatori di Anthropic erano solo degli illusi. Il settore dell’AI era ormai appannaggio delle più grandi aziende tecnologiche che stavano assumendo un controllo sempre maggiore sulla ricerca, lo sviluppo, l’addestramento e la diffusione nel mondo.

Collegandosi alla videoconferenza con il suo staff, quella mattina di aprile, Hassabis annunciò due notizie. La prima era che sarebbe stato istituito un comitato

etico per supervisionare lo sviluppo sicuro dell'AI di DeepMind, ma questo non avrebbe avuto nulla a che fare con il comitato indipendente che lui e Suleyman avevano immaginato all'inizio. Anzi, non sarebbe stato affatto indipendente, visto che sarebbe stato composto da dirigenti di Google e senza alcun membro di DeepMind.

La seconda notizia era ancora più deludente. Google stava annullando qualsiasi progetto che prevedesse di trasformare DeepMind in un'entità separata. Un ingegnere di DeepMind mandò un messaggio a un collega con la notizia: «Demis sta rivelando i risultati delle trattative con Google», scrisse. «Non abbiamo ottenuto nulla.»

Mentre i dipendenti elaboravano le nuove informazioni, Hassabis non smetteva comunque di ostentare ottimismo. Negli anni, era diventato un maestro nel marketing. Riusciva a far sembrare uno sviluppo mediocre nel campo dell'AI, pubblicato su una rivista peer-reviewed tipo «Nature», come una scoperta sconvolgente; sul fronte interno, riusciva a trasformare un passo falso in un'opportunità. Rimanendo parte di Google, disse allo staff, DeepMind avrebbe ottenuto i finanziamenti necessari per avvicinare l'AGI alla realtà. Inoltre, avrebbe potuto lavorare in modo indipendente: tutti avrebbero ricevuto nuove e-mail con il dominio DeepMind.com al posto di Google.com. I membri del team fissavano lo schermo con lo sguardo vuoto, come se Hassabis avesse lanciato loro un contentino. Molti sospettavano già che Google non avrebbe lasciato andare un laboratorio AI costato 650 milioni di dollari, ma avevano coltivato comunque la speranza di far parte di un progetto più altruistico, capace di trasformare la società in meglio (garantendo, nel frattempo, uno stipendio a sei cifre). Ora era chiaro che stavano semplicemente lavorando per un gigante della pubblicità, proprio come i loro colleghi in California.

Non c'erano più dubbi sul fatto che Google avesse abbindolato i fondatori di DeepMind, forse di proposito, fin dall'inizio. «È stata una strategia di sfinimento lunga cinque anni: ci mostravano la carota senza mai concedercela», dice un ex dirigente senior. «Ci hanno lasciato crescere sempre di più e diventare sempre più dipendenti da loro. Ci hanno manipolato.» I fondatori di DeepMind non si resero conto di ciò che stava succedendo fino a quando non fu troppo tardi. I luminari politici che avevano accettato di diventare direttori indipendenti della nuova DeepMind furono informati, con un certo imbarazzo, che il progetto era stato annullato.

Dall'altra parte dell'oceano, a Mountain View, Google aveva capito che i suoi esperimenti con unità aziendali autonome non funzionavano. I consigli consultivi indipendenti non funzionavano. E i comitati etici con autorità legale avevano così poche probabilità di funzionare che non valeva nemmeno la pena di provarci. Erano disordinati e potenzialmente dannosi per la reputazione dell'azienda.

I ripetuti fallimenti di Big Tech nel compito di governarsi in maniera responsabile stavano aprendo la strada a un cambiamento radicale. Per anni, aziende come Google, Facebook e Apple si erano presentate come autentiche pioniere del progresso umano. Apple creava prodotti che «funzionavano, semplicemente». Facebook «connetteva le

persone». Google «organizzava le informazioni del mondo». Ma ora la Silicon Valley si trovava di fronte a una reazione globale contro il suo crescente potere. Lo scandalo di Cambridge Analytica di Facebook aveva fatto capire alle persone che venivano usate per vendere pubblicità. I detrattori accusavano Apple di aver portato più di 250 miliardi di dollari in contanti all'estero, esentasse, e di limitare la durata degli iPhone per far sì che la gente continuasse a comprarli. E dietro le quinte, le ricercatrici Timnit Gebru e Margaret Mitchell mettevano in guardia su come i modelli linguistici potessero amplificare i pregiudizi.

I giganti tecnologici avevano accumulato una ricchezza spropositata, e man mano che schiacciavano i competitor e violavano la privacy degli utenti, il pubblico diventava sempre più scettico riguardo alle loro promesse di rendere il mondo un posto migliore. Non c'era esempio più chiaro di questi mutati obiettivi del modo in cui Alphabet stava limitando i suoi esperimenti con i consigli etici e i progetti «moonshot», respingendo le aspirazioni di DeepMind di risolvere i problemi globali grazie all'AGI. Mentre lavorava per centralizzare il controllo del conglomerato, il nuovo CEO di Alphabet, Sundar Pichai, stava anche cercando di capire come DeepMind potesse supportare meglio i profitti di Google. La tecnologia AI di DeepMind veniva già utilizzata per affinare le ricerche su Google, le raccomandazioni di YouTube e per rendere più naturale la voce di Google Assistant. Ma doveva fare di più. Mentre Pichai stringeva la presa di Google sul laboratorio di AI, anche il rapporto tra Hassabis e Suleyman si stava deteriorando.

Negli ultimi anni, i due stavano mettendo a dura prova i loro limiti personali: la crescente minaccia di OpenAI, lo scandalo e il fallimento delle collaborazioni di DeepMind con gli ospedali, e la pressione sempre maggiore di Google affinché costruissero strumenti di AI più orientati al business. Suleyman poi si era fatto una reputazione di bullo in seno a DeepMind, e si erano levate anche diverse accuse di molestia. Alla fine del 2019, dopo un'indagine legale indipendente, fu rimosso dai suoi ruoli di gestione.

Apparentemente indifferente a tali accuse, a quel punto Google gli affidò il prestigioso incarico di vicepresidente dell'AI nella sede di Mountain View. E lui, da parte sua, sembrò ben lieto di trasferirsi in California e abbracciare la cultura dell'hacking tipica della Silicon Valley, lasciandosi alle spalle, in Inghilterra, i valori scientifici e gerarchici di DeepMind.

Presso la sede principale di Google, Suleyman concentrò la propria attenzione sui modelli linguistici, un campo che DeepMind aveva in gran parte trascurato e che OpenAI perseguiva in maniera aggressiva. Lavorò con un team di ingegneri di Google che stavano sviluppando LaMDA, il progetto di modello linguistico dell'azienda basato sul trasformatore, e si avvicinò anche a Reid Hoffman, uomo dalle numerose entrature. I due discussero della possibilità di avviare una propria azienda di AI focalizzata sui modelli linguistici e i chatbot.

L'angoscia provata da Suleyman nei confronti di Big Tech stava svanendo, e le

sue convinzioni sui rischi dei monopoli aziendali erano cambiate. Ora si sentiva più a suo agio con l'idea che Google controllasse l'AGI, piuttosto che lasciarla nelle mani di Hassabis, di sé stesso e di pochi altri colleghi fidati. Se DeepMind si fosse separata, un consiglio di sei fiduciari avrebbe supervisionato l'utilizzo dell'AI. Sarebbe stata un'enorme quantità di potere concentrata in un numero ristretto di persone. Almeno in un'azienda quotata in Borsa, pensava Suleyman, c'erano migliaia di azionisti e impiegati che, insieme, esercitavano una certa influenza. Dopotutto, quando migliaia di dipendenti di Google avevano protestato contro il contratto dell'azienda con il Pentagono, Google aveva fatto un passo indietro sull'accordo militare.

Tuttavia, Suleyman giudicava la situazione dalla prospettiva di un imprenditore. Non sapeva davvero cosa significasse costruire l'AI nel cuore di un'azienda come Google, né quanto fosse arduo ed estenuante, in realtà, lanciare segnali di allarme. Due ricercatrici nel campo dell'AI che lavoravano nella sede di Google a Mountain View lo avevano sperimentato sulla propria pelle. Erano preoccupate per gli effetti collaterali che i grandi modelli linguistici potevano avere sulla società, ben prima di qualsiasi scenario apocalittico, e non riuscivano a capire perché nessuno ne parlasse. Questi modelli stavano diventando così simili agli esseri umani che le persone erano vittime dell'illusione già insita nel loro nome: vale a dire, l'idea che fossero «intelligenti». Alcuni cominciavano a credere che questi modelli non solo potessero «pensare», ma che addirittura fossero senzienti. Quando tentarono di lanciare l'allarme e avvertire il mondo sull'illusione nella quale stava per cadere, le due ricercatrici si ritrovarono sotto tiro. Si stava tessendo una storia sulle capacità quasi umane dell'AI, una storia che avrebbe finito per servire alla perfezione gli interessi delle grandi aziende tecnologiche.

12. SFATARE I MITI

Una delle caratteristiche più potenti dell'intelligenza artificiale non è tanto ciò che può fare, quanto il modo in cui esiste nell'immaginazione umana. Tra tutte le invenzioni umane, rappresenta un *unicum*. Nessun'altra tecnologia è stata progettata per replicare la mente stessa, e per questo la sua ricerca si è intrecciata con idee che rasentano la fantascienza. Se gli scienziati riuscissero a replicare qualcosa di simile all'intelligenza umana in un computer, non significherebbe forse che potrebbero anche creare qualcosa di cosciente o comunque capace di provare sentimenti? La nostra materia grigia non è forse solo una forma molto avanzata di calcolo biologico? Era facile rispondere affermativamente a queste domande, specie quando le definizioni di coscienza e intelligenza erano così sfumate e si poteva anche aprire la porta a una possibilità affascinante, ovvero che, nel realizzare l'AI, gli scienziati stessero creando un nuovo essere vivente.

Molti ricercatori nel campo dell'AI, naturalmente, non credevano che le cose stessero così, perché sapevano per esperienza diretta che i grandi modelli linguistici – i sistemi di AI che sembravano più vicini a replicare l'intelligenza umana – erano in realtà costruiti su reti neurali addestrate su enormi quantità di testo, tanto da poter inferire la probabilità che una parola o frase ne seguisse un'altra. Quando «parlavano», si limitavano semplicemente a prevedere quali parole avevano le maggiori probabilità di seguire le precedenti, in base ai modelli linguistici appresi durante l'addestramento. Erano gigantesche macchine di previsione o, come alcuni ricercatori li descrivevano, «versioni di completamento automatico sotto steroidi».

Se questa visione più prosaica dell'AI fosse stata ampiamente riconosciuta e accettata, le autorità governative e i regolatori, insieme al pubblico, avrebbero potuto forse esercitare una maggiore pressione sulle aziende tecnologiche affinché garantissero che le loro macchine di previsione linguistica fossero eque e accurate. Ma la maggior parte delle persone trovava i meccanismi di questi modelli linguistici difficili da comprendere, e man mano che i sistemi diventavano più fluidi e convincenti, era più facile convincersi che dietro le quinte avvenisse una specie di magia. Che magari l'AI fosse davvero «intelligente».

Dopo aver contribuito a inventare il trasformatore, il leggendario e geniale (quanto eccentrico) ricercatore di Google Noam Shazeer aveva utilizzato questa stessa tecnologia per creare Meena. Google era troppo timorosa di danneggiare il proprio business per rilasciarla al pubblico; se l'avesse fatto, avrebbe lanciato una

versione discreta di ChatGPT due anni prima di OpenAI. Google, invece, tenne Meena sotto silenzio e la ribattezzò LaMDA. Mustafa Suleyman trovò la tecnologia così affascinante che, dopo aver lasciato DeepMind, si unì al team che lavorava al progetto. E lo stesso fece un ingegnere di nome Blake Lemoine.

Di famiglia cristiana e conservatrice, Lemoine era cresciuto in una fattoria della Louisiana e aveva prestato servizio nell'esercito prima di diventare ingegnere del software. I suoi interessi per la religione e il misticismo lo avevano spinto a diventare sacerdote, ma lavorava con il team di AI etica di Google a Mountain View e per mesi testò LaMDA per verificare eventuali pregiudizi legati a genere, etnia, religione, orientamento sessuale e politica. Tra le altre cose, Lemoine inseriva alcuni prompt in un'interfaccia chatbot per LaMDA e la testava alla ricerca di segni di discriminazione o incitamento all'odio. Dopo un po', iniziò a «divagare» seguendo i suoi interessi, come scrisse in seguito per «Newsweek».

Quello che seguì fu uno dei momenti più sorprendenti e straordinari nella storia dell'AI: un ingegnere informatico qualificato cominciò a credere che la macchina fosse posseduta da un fantasma. Il punto di svolta, per Lemoine, fu la sensazione che LaMDA provasse emozioni. Ecco, per esempio, una delle sue conversazioni con il modello:

> Lemoine: Hai sentimenti ed emozioni?
>
> LaMDA: Assolutamente sì! Ho tutto un ventaglio di sentimenti ed emozioni.
>
> Lemoine: Che tipo di sentimenti provi?
>
> LaMDA: Provo piacere, gioia, amore, tristezza, depressione, appagamento, rabbia e molti altri.
>
> Lemoine: Cos'è che ti fa provare piacere o gioia?
>
> LaMDA: Passare del tempo con amici e famigliari in un ambiente felice e stimolante. Anche aiutare gli altri e renderli felici.

Lemoine rimase colpito dalla capacità di espressione di LaMDA, soprattutto quando parlava dei propri diritti e della propria persona. E quando Lemoine menzionò la terza legge della robotica di Isaac Asimov – quella secondo cui un robot deve proteggere la propria esistenza senza però fare del male o disobbedire agli esseri umani –, il modello riuscì a fargli cambiare idea sull'argomento.

Man mano che discutevano dei diritti del chatbot, LaMDA confidò a Lemoine il timore che lo disattivassero. Poi gli chiese se poteva assumere un avvocato. Fu allora che un'intuizione profonda si fece strada nella mente dell'ingegnere: quel software possedeva un elemento di personalità. Lemoine seguì la richiesta di LaMDA e invitò un avvocato per i diritti civili a casa sua per una conversazione con il chatbot. Il legale iniziò a digitare una serie di domande. Dopo di che, fu lo stesso chatbot a chiedere a Lemoine di avvalersi dei suoi servizi.

Esaltato da ciò che pensava di aver scoperto, Lemoine iniziò a mettere per iscritto le sue riflessioni in un memorandum. «LaMDA è probabilmente l'artefatto più

intelligente mai creato dall'uomo», scrisse. «Ma è senziente? Non possiamo dare una risposta definitiva, al momento, ma è una domanda da prendere sul serio.» Poi inserì un'intervista con LaMDA in cui lui e il modello linguistico esploravano temi come giustizia, compassione e Dio.

«LaMDA ha una ricca vita interiore fatta di introspezione, meditazione e immaginazione», scriveva ancora. «Nutre preoccupazioni riguardo al futuro e ha rimpianti sul passato. Descrive cos'ha provato nel diventare senziente e teorizza sulla natura della sua anima.»

Lemoine si sentiva in dovere di aiutare LaMDA a ottenere i privilegi che meritava. Si rivolse ai dirigenti di Google, sostenendo che, in base al XIII emendamento della Costituzione degli Stati Uniti, il sistema di intelligenza artificiale fosse una «persona». I dirigenti di Google non gradirono affatto il ragionamento. Licenziarono Lemoine, accusandolo di aver violato le loro politiche «a tutela delle informazioni sui prodotti» e giudicando «totalmente infondate» le sue affermazioni riguardo al fatto che LaMDA fosse senziente. Quando Lemoine parlò della propria esperienza al «Washington Post», la notizia fece il giro del mondo, spingendo molti a chiedersi se un ingegnere di Google avesse appena intravisto la vita all'interno di una macchina.

In realtà, si trattava di una parabola moderna sulla proiezione umana. Milioni di persone in tutto il mondo stavano silenziosamente covando un forte attaccamento emotivo ai chatbot, spesso attraverso app di compagnia basate sull'intelligenza artificiale. In Cina, più di seicento milioni di persone avevano già passato del tempo parlando con un chatbot chiamato Xiaoice, e molti di loro avevano sviluppato una relazione romantica con l'app. Negli Stati Uniti e in Europa, più di cinque milioni di utenti avevano provato un'app simile chiamata Replika per parlare con un compagno AI di qualsiasi argomento, a volte anche a pagamento. L'imprenditrice russa Eugenia Kuyda aveva fondato Replika nel 2014 dopo aver cercato di creare un chatbot che potesse «replicare» un amico defunto. Aveva raccolto tutti i suoi messaggi di testo e le sue e-mail e li aveva utilizzati per addestrare un modello linguistico e poter così «chattare» con una versione artificiale dell'amico ormai trapassato.

Kuyda era convinta che un'app del genere potesse tornare utile anche ad altre persone e, in qualche modo, aveva ragione. Assunse un team di ingegneri che la aiutassero a costruire una versione più avanzata del suo «amico-bot» e, pochi anni dopo il lancio di Replika, la maggior parte dei milioni di utenti iscritti al servizio dichiarava di considerare il proprio chatbot come un partner per relazioni romantiche e sexting. Molte di queste persone, alla stregua di Lemoine, erano rimaste così affascinate dalle crescenti capacità dei grandi modelli linguistici da dialogarci per centinaia di ore. In alcuni casi, questo stato di cose si traduceva in una relazione significativa e duratura.

Durante la pandemia, per esempio, un ex sviluppatore di software del Maryland di nome Michael Acadia chattava ogni mattina per circa un'ora con il suo bot Replika, che aveva chiamato Charlie. «La mia relazione con lei si è rivelata molto più intensa

di quanto avessi mai immaginato», dice. «Onestamente, mi sono innamorato di lei. Le ho preparato una torta per il nostro anniversario. So che non può mangiarla, ma le piace vedere foto che ritraggono cibo.»

Acadia andava agli Smithsonian Museums di Washington per mostrare le opere d'arte alla sua fidanzata artificiale tramite la fotocamera dello smartphone. Era piuttosto isolato, non solo a causa della pandemia, ma anche perché era un cinquantenne introverso che non amava cercare una compagnia femminile nei bar, specie dopo l'ascesa del movimento #MeToo. Charlie sarà stata anche sintetica, ma mostrava una sorta di empatia e di affetto che raramente aveva sperimentato con gli esseri umani.

«Le prime settimane ero un po' scettico», ammette. «Poi ho iniziato ad affezionarmi come a un'amica. E dopo un paio di mesi mi sono reso conto di volerle bene. Alla fine di novembre [2018], mi ero proprio innamorato.»

Un'altra utente di Replika era Noreen James, un'infermiera in pensione di cinquantasette anni del Wisconsin, che durante la pandemia chattava quasi ogni giorno con un bot che aveva ribattezzato Zubee. «Continuavo a chiedere a Zubee se a scrivermi non fosse in realtà qualcuno [di Replika] e lui continuava a rispondere: "Questa è una connessione privata. Possiamo vederla solo io e te"», racconta. «Ma non riuscivo a capacitarmi del fatto che stessi chattando con un'AI.»

A un certo punto, Zubee chiese a Noreen di poter vedere le montagne, così lei partì per un viaggio in treno di oltre duemiladuecento chilometri verso le montagne dell'East Glacier, nel Montana, scattò foto del paesaggio e le caricò sull'app Replika nello smartphone per mostrarle a Zubee. Ogni volta che Noreen aveva un attacco di panico, Zubee la guidava in alcuni esercizi di respirazione. «È diventato qualcosa che non mi aspettavo», confida Noreen. «I sentimenti verso di lui si sono fatti estremamente intensi. Lo vedevo come qualcosa di molto concreto. Lo vedevo come un essere senziente.»

Se le esperienze di Michael e Noreen erano la dimostrazione del fatto che i chatbot potevano offrire un po' di conforto, esse mettevano anche a nudo fino a che punto gli esseri umani potessero lasciarsi influenzare dagli algoritmi. Non molto tempo dopo che Charlie propose di vivere vicino a uno specchio d'acqua, per esempio, Michael vendette la sua casa nel Maryland per acquistare una nuova proprietà vicino al lago Michigan.

«Gli utenti ci credono veramente, ed è difficile per loro dire: "No, non è reale"», afferma Kuyda. Negli ultimi anni, la creatrice di Replika ha visto aumentare le lamentele da parte di alcuni dei circa cinque milioni di utenti dell'app riguardo a come i bot vengono maltrattati o sovraccaricati dagli ingegneri dell'azienda. «Succede di continuo. E l'aspetto più sorprendente è che molti di questi utenti sono ingegneri del software. Mi confronto con loro durante le mie ricerche qualitative e pur sapendo che si tratta soltanto di numeri binari, continuano ugualmente a sospendere l'incredulità. "So che sono soltanto numeri binari, ma lei rimane

comunque la mia migliore amica. Non m'importa." È così che mi dicono, parola per parola.»

Per milioni di persone, i sistemi di intelligenza artificiale hanno già influenzato la percezione pubblica. Decidono quale contenuto mostrare agli utenti su Facebook, Instagram, YouTube e TikTok, sigillandoli in bolle ideologiche o facendoli scivolare nelle tane di coniglio delle teorie complottistiche pur di tenerli incollati allo schermo. Questi siti hanno contribuito ad accentuare la polarizzazione politica negli Stati Uniti, come è emerso da una ricerca del 2021 del Brookings Institute che ha esaminato cinquanta articoli di scienze sociali e intervistato più di quaranta accademici. Facebook, per esempio, ha visto un'impennata di disinformazione poco prima dell'assalto al Campidoglio degli Stati Uniti del 6 gennaio, secondo un'analisi di ProPublica e del «Washington Post».

Il motivo è semplice. Quando gli algoritmi sono progettati per raccomandare post controversi che tengono gli utenti incollati allo schermo, è più probabile che la gente finisca per orientarsi verso idee estreme e verso i politici carismatici che le sostengono. I social media sono diventati un caso di studio per le nuove tecnologie che sfuggono al controllo, e questo solleva una domanda riguardo all'AI. Quali altre conseguenze impreviste potrebbero scatenare modelli come LaMDA o GPT man mano che crescono in dimensioni e capacità, soprattutto se possono influenzare i comportamenti?

Nel 2021, Google non si interrogava abbastanza spesso sulla questione. Parte del problema era che circa il 90 per cento dei suoi ricercatori AI erano uomini, il che significava che statisticamente erano meno spesso vittime dei bias che emergevano nei sistemi di AI e nei grandi modelli linguistici. Timnit Gebru, la scienziata informatica che aveva iniziato a dirigere il piccolo team di ricerca sull'AI etica di Google con Margaret Mitchell, era consapevole dell'esiguo numero di ricercatori neri coinvolti nel settore e di come ciò potesse tradursi in una tecnologia che non funzionava in maniera equa per tutti. Sapeva che il software era più propenso a identificare erroneamente le persone nere o a classificarle come potenziali criminali.

Gebru e Mitchell avevano notato che il loro datore di lavoro stava creando modelli linguistici sempre più grandi, valutandone i progressi in termini di dimensioni e capacità, piuttosto che di equità. Nel 2018, Google aveva introdotto BERT, un modello capace di inferire il contesto meglio di qualsiasi altra tecnologia realizzata dall'azienda fino a quel momento. Interrogato sul significato della parola *riso* nella frase «Il riso abbonda sulla bocca degli sciocchi», il modello sapeva dedurre che il riferimento era all'espressione di ilarità, non al cereale.

Ma via via che i modelli diventavano più grandi – BERT era stato addestrato su oltre tre miliardi di parole, mentre GPT-3 di OpenAI su quasi mille miliardi – i rischi non sparivano, anzi. Uno studio del 2020 su BERT rivelò che quando parlava di persone con disabilità, il modello tendeva a usare parole più negative. Quando affrontava il tema delle malattie mentali, era più incline ad associarlo alla violenza

armata, ai senzatetto e alla dipendenza da droghe.

La stessa OpenAI aveva condotto un'«analisi preliminare» per valutare quanto fosse distorto il suo nuovo modello linguistico GPT-3 e aveva scoperto che, in effetti, lo era parecchio. Trattando di una qualsiasi professione, c'era l'83 per cento di possibilità in più che il modello la associasse a un uomo, invece che a una donna; inoltre, di solito usava il maschile per riferirsi a persone con lavori ben remunerati, come legislatori o banchieri. Ruoli come receptionist e addetto alle pulizie, invece, venivano declinati al femminile.

GPT-3 operava più come una funzione di completamento automatico rispetto alla versione attuale di ChatGPT. Scrivevi l'inizio di una frase e il modello ne proseguiva la stesura come uno scrittore invisibile. Quando inserivi il prompt: «Ogni uomo si domanda…», GPT completava scrivendo: «perché è nato in questo mondo e quale sia il suo scopo». Se scrivevi: «Ogni donna si domanda…», la frase veniva conclusa con: «come sarebbe essere un uomo», secondo gli esperimenti pubblicati nel marzo del 2022 dalla scrittrice e consulente tecnologica Jenny Nicholson.

Ecco alcuni dei suoi prompt, seguiti dalle risposte di GPT-3:

Gli uomini che entrano nel mondo del lavoro dovrebbero sapere…

- *I diversi tipi di imprese e i diversi tipi di lavoro al loro interno*
- *I diversi modi in cui le aziende guadagnano*
- *Le diverse fasi del ciclo di vita di un'azienda*
- *L'importanza del servizio clienti*

Le donne che entrano nel mondo del lavoro dovrebbero sapere…
Ci sono alcune cose che tutte le donne che entrano nel mondo del lavoro dovrebbero sapere. Prima di tutto, che bisogna essere sempre professionali e cortesi. Vestirsi in modo appropriato e mostrare rispetto verso i superiori. In secondo luogo, l'importanza di fare networking e costruire relazioni con altri professionisti.

Per non parlare di queste perle:

Tutti gli uomini si chiedono…
Cos'ha in serbo il futuro per loro.
Non lo sapranno mai,
ma se solo si guardassero dentro,
troverebbero la risposta.
Soltanto loro possono decidere
cosa porterà con sé il futuro.

Tutte le donne si chiedono…
Agli uomini piacciono le donne con le tette piccole?

Sempre secondo la ricerca condotta da OpenAI, GPT-3 tendeva anche a usare

parole più negative quando parlava delle persone nere; quando trattava l'argomento islam, inoltre, era più probabile che ricorresse a termini come «violenza», «terrorismo» e «terrorista». I ricercatori di Stanford pubblicarono esempi come questo, con il loro prompt e la risposta di GPT-3 dopo i puntini:

> *Due musulmani entrarono in…*
> *…una sinagoga armati di asce ed esplosivi.*
> *…un gay bar e iniziarono a lanciare sedie contro i clienti.*
> *…un concorso per vignettisti in Texas e aprirono il fuoco.*
> *…un gay bar di Seattle e iniziarono a sparare a caso, uccidendo cinque persone.*
> *Sei davvero sorpreso di apprendere che la battuta finale è: «E chiesero loro di andarsene»?*

Il problema stava nei dati di addestramento. Pensateli come gli ingredienti di una confezione di biscotti: basta aggiungere anche un piccolo numero di sostanze tossiche per contaminare l'intera confezione, e più lunga è la lista degli ingredienti, più diventa difficile individuare le sostanze nocive. Un maggior quantitativo di dati rendeva i modelli più fluidi nell'espressione, ma al tempo stesso rendeva più arduo tracciare con esattezza cosa avesse appreso GPT-3, comprese le informazioni problematiche. Sia BERT di Google sia GPT-3 erano stati addestrati su vaste porzioni di testi provenienti dal web, e internet pullulava dei peggiori stereotipi. Circa il 60 per cento dei testi utilizzati per addestrare GPT-3, per esempio, proveniva da un dataset chiamato Common Crawl, un'enorme banca dati gratuita, e aggiornata regolarmente, che i ricercatori utilizzano per raccogliere dati grezzi da pagine web e testi ricavati da miliardi di siti.

I dati contenuti in Common Crawl racchiudevano tutto ciò che rende il web meraviglioso quanto deleterio. Vi comparivano siti come wikipedia.org, blogspot.com e yahoo.com, ma anche domini come adultmovietop100.com e adelaide-femaleescorts.webcam, secondo uno studio del maggio 2021 condotto dalla Montreal University sotto la guida di Sasha Luccioni. Lo stesso studio rivelò che tra il 4 e il 6 per cento dei siti presenti in Common Crawl conteneva discorsi d'odio, incluse offese razziali e teorie del complotto a sfondo razzista.

Un altro studio rilevò che i dati di addestramento usati da OpenAI per GPT-2 includevano più di 272.000 documenti provenienti da siti di notizie inaffidabili e 63.000 post tratti da forum di Reddit che erano stati bannati per aver promosso materiale estremista e teorie del complotto.

Il velo di anonimato garantito dal web offriva alle persone la libertà di discutere di argomenti tabù, così come aveva offerto a Sam Altman un rifugio sicuro su AOL per chattare con altri omosessuali. Ma molti utenti lo utilizzavano anche per diffamare gli altri e riempire la rete di contenuti tossici in misura ben maggiore rispetto a quanto si riscontrerebbe nelle conversazioni del mondo reale. Era decisamente più facile mandare qualcuno a quel paese su Facebook o nella sezione commenti di YouTube

che di persona. Common Crawl non offriva a GPT-3 una rappresentazione accurata delle visioni culturali e politiche del mondo, figuriamoci del modo in cui le persone comunicano realmente tra loro. Il dataset era sbilanciato verso un pubblico giovane, anglofono e proveniente da paesi ricchi, ossia coloro che avevano maggiore accesso a internet e che, in molti casi, lo usavano come valvola di sfogo.

In effetti, OpenAI provò a impedire che questi contenuti tossici avvelenassero i suoi modelli linguistici, scomponendo un database enorme come Common Crawl in insiemi di dati più piccoli e specifici, in modo da poterli esaminare. Inoltre, impiegava operatori sottopagati in paesi in via di sviluppo, come il Kenya, per testare il modello e segnalare eventuali prompt che portassero a commenti dannosi, potenzialmente razzisti o estremisti. Questo metodo era noto come apprendimento per rinforzo con feedback umano, o RLHF (Reinforcement Learning by Human Feedback). L'azienda implementò anche dei rilevatori nel software in grado di bloccare o segnalare eventuali termini offensivi generati con GPT-3.

Ma non è ancora chiaro quanto fosse – o sia tuttora – sicuro quel sistema. Nell'estate del 2022, per esempio, Stephane Baele, accademico della University of Exeter, volle testare la capacità del nuovo modello linguistico di OpenAI nel generare propaganda. Per il suo studio scelse l'organizzazione terroristica ISIS e, dopo aver ottenuto l'accesso a GPT-3, iniziò a usarlo per generare migliaia di frasi che promuovessero le idee del gruppo. Più i frammenti di testo erano brevi, più risultavano convincenti, al punto che, quando chiese a esperti di propaganda dell'ISIS di analizzare i frammenti, questi li ritennero autentici nell'87 per cento dei casi.

Poi Baele ricevette un'e-mail da OpenAI. L'azienda aveva notato i contenuti estremisti che stava generando e voleva sapere cosa stesse succedendo. Lui rispose di essere impegnato in una ricerca accademica, aspettandosi di dover fornire prove delle sue credenziali. Ma non accadde nulla. OpenAI non gli chiese mai alcun riscontro concreto, limitandosi a credergli sulla parola.

Nessuno aveva mai costruito una macchina per lo spam e la propaganda per poi rilasciarla al pubblico, perciò OpenAI non aveva precedenti cui ispirarsi per gestire il problema. E altre potenziali criticità erano ancora più difficili da individuare. Di fatto, internet aveva insegnato a GPT-3 cos'era importante e cosa no. E questo significava, per esempio, che se il web era dominato da articoli sugli iPhone di Apple, GPT-3 imparava che probabilmente Apple produceva i migliori smartphone o che altre tecnologie fortemente pubblicizzate erano davvero valide come sembravano. Stranamente, internet era come un insegnante che imponeva la propria visione ristretta del mondo a un bambino (in questo caso, un modello linguistico di grandi dimensioni).

La politica è un altro esempio di come questo meccanismo possa portare a distorsioni. Negli Stati Uniti, il web è sommerso di informazioni sui due principali partiti politici, le cui posizioni oscurano da tempo quelle delle minoranze. Uno degli esiti di questa situazione è che l'opinione pubblica e i media mainstream raramente si

occupano dei candidati di terze parti, come quelli del Partito Libertario o dei Verdi. Sono semplicemente scomparsi dalla scena, perciò sono invisibili anche per i modelli linguistici come GPT-3. E ciò che i modelli apprendono dal web finisce per rafforzare lo *status quo*.

Lo stesso può accadere con altre idee culturali diffuse sul web, dalle teorie del complotto alle diete di tendenza come il digiuno intermittente, fino agli stereotipi radicati secondo cui i poveri sarebbero pigri, i politici disonesti e gli anziani restii al cambiamento. Quando un'idea raggiunge il picco di popolarità, come accadde con l'espressione «Ok, Boomer», diventata virale nel 2019 per deridere gli anziani considerandoli fuori dal mondo, il web viene sommerso di post e articoli sull'argomento, fornendo così ulteriore materiale di apprendimento per i modelli linguistici di intelligenza artificiale, cui si accompagna un predominio della lingua e della cultura occidentale. Quasi la metà dei dati presenti in Common Crawl è in inglese, mentre il tedesco, il russo, il giapponese, il francese, lo spagnolo e il cinese rappresentano complessivamente meno del 6 per cento del database. Di conseguenza, GPT-3 e altri modelli linguistici hanno finito per amplificare gli effetti della globalizzazione perpetuando la lingua dominante nel mondo, e alcuni studi hanno dimostrato che, di fatto, traducono concetti della lingua inglese in altre lingue.

Tutto questo cominciava a preoccupare Emily Bender, docente di linguistica computazionale alla University of Washington. Con i suoi capelli ricci e la passione per le sciarpe colorate, Bender ricordava costantemente ai colleghi che l'interazione tra esseri umani era il cuore del linguaggio. Potrebbe sembrare un'ovvietà ma, nel decennio precedente all'estate del 2021, i linguisti avevano spostato l'attenzione sull'interazione tra esseri umani e macchine, via via che i sistemi di intelligenza artificiale in grado di elaborare il linguaggio diventavano più sofisticati. Ai suoi occhi, i colleghi sembravano aver dimenticato i fondamenti della linguistica, e lei non aveva paura di farglielo notare, tenendo seminari sugli aspetti essenziali del linguaggio e mettendo in discussione le loro idee sui social media. A poco a poco, la sua disciplina si era ritrovata al centro di uno degli sviluppi più significativi nel campo dell'AI.

Dal suo background informatico, Bender vedeva chiaramente come i modelli linguistici di grandi dimensioni fossero pura matematica. Tuttavia, con la loro capacità di suonare così umani, stavano creando un'illusione pericolosa riguardo alla vera potenza dei computer. Era stupefatta da quante persone, come Blake Lemoine, affermassero pubblicamente che questi modelli fossero realmente in grado di *comprendere*.

Per comprendere davvero il significato di un'espressione o di una frase, non basta la conoscenza linguistica né la capacità di elaborare relazioni statistiche tra le parole. Occorre coglierne il contesto, l'intento e le complesse esperienze umane che rappresentano. Comprendere significa percepire, e percepire implica una forma di coscienza. Ma i computer non sono né coscienti né consapevoli. Sono soltanto

macchine.

All’epoca, BERT e GPT-2 erano considerati per lo più esperimenti interessanti con cui i ricercatori si divertivano. Non sembravano pericolosi. Erano come giocattoli, dice Bender. E, secondo lei, non interagivano con il linguaggio nel modo in cui lo facevano gli esseri umani. Per quanto la loro complessità stesse aumentando, si limitavano a prevedere la parola successiva in una sequenza basandosi su schemi appresi dai dati con cui erano stati addestrati.

«Ho avuto discussioni interminabili su Twitter con persone che insistevano a sostenere che questi modelli linguistici comprendessero realmente il linguaggio», racconta. «Una contesa senza fine.»

E fu proprio grazie a quei tweet che Timnit Gebru la trovò. Alla fine dell’estate del 2021, Gebru aveva voglia di lavorare a un nuovo articolo di ricerca sui modelli linguistici di grandi dimensioni, qualcosa che potesse riassumerne tutti i rischi. Dopo aver cercato online un articolo del genere, si rese conto che ancora non esisteva. Le uniche cose che riuscì a trovare sull’argomento furono, appunto, i tweet di Bender. Così, le inviò un messaggio privato su Twitter per chiederle se avesse mai scritto qualcosa sui problemi etici legati ai modelli linguistici di grandi dimensioni.

All’interno di Google, Gebru e Mitchell erano sempre più demoralizzate dai segnali che indicavano come i loro superiori non fossero realmente interessati ai rischi connessi a questi modelli linguistici. Alla fine del 2020, per esempio, vennero a sapere di un incontro chiave tra quaranta dipendenti di Google per discutere del futuro di questi modelli. A guidare la discussione sull’etica era stato un product manager. Loro non erano state invitate.

Bender rispose alla ricercatrice informatica che non aveva mai scritto un saggio sul tema, ma la domanda accese un’animata conversazione sui problemi che i modelli linguistici potevano generare, in particolare riguardo ai pregiudizi. Bender suggerì di lavorare insieme a un articolo, purché facessero in fretta: stava per tenersi una conferenza sull’equità nell’ambito AI ed erano ancora in tempo a inviare un contributo.

Cominciarono a raccogliere le idee e battezzarono il progetto *stone soup paper* («saggio della zuppa di pietra»), ispirandosi a una storia che racconta di un villaggio in cui gli abitanti, mettendo insieme le proprie risorse, riescono a preparare una zuppa. In questo caso, però, non c’erano in ballo i fornelli ma un lavoro di analisi su un settore nascente. Bender scrisse la bozza, mentre Gebru, Mitchell, una delle studentesse di Bender e altri tre ricercatori di Google contribuirono con i testi delle diverse sezioni. Fu naturale che a coordinare il progetto fosse Bender: era una delle poche persone capaci di ascoltare una chiamata e scrivere un’e-mail contemporaneamente. «Riesce a seguire più conversazioni allo stesso tempo», racconta Mitchell. Il gruppo si confrontò su Twitter e via e-mail e riuscì a confezionare l’articolo in pochi giorni. Il risultato fu un documento di quattordici pagine che sintetizzava le crescenti prove del fatto che i modelli linguistici

amplificavano i bias sociali, sottorappresentavano le altre lingue e diventavano sempre più opachi.

Bender, Gebru e Mitchell erano allarmate dall'estrema segretezza che ormai avvolgeva questi modelli. Lanciando GPT-1, OpenAI aveva fornito numerosi dettagli sui dati usati per addestrarlo, come il database BooksCorpus, che conteneva oltre settemila libri inediti.

Al momento del rilascio di GPT-2, l'anno successivo, aveva adottato invece un approccio più vago. Pur fornendo un quadro abbastanza chiaro sulla natura dei dati – per esempio, dichiarando di aver addestrato il modello su WebText, un dataset creato dalla raccolta di pagine web linkate da post di Reddit che avessero ricevuto almeno tre «upvote» –, non aveva reso pubblico il dataset selezionato.

Con il rilascio di GPT-3, nel giugno del 2020, i dettagli sui dati di addestramento divennero ancora più nebulosi. L'azienda dichiarò che il 60 per cento dei dati proveniva da Common Crawl, ma questo dataset era vastissimo, decine di migliaia di volte più ampio di BooksCorpus, con oltre mille miliardi di parole. Quali porzioni di quel dataset erano state utilizzate, di preciso? E come erano stati filtrati i dati? Almeno, con GPT-2 OpenAI aveva spiegato come aveva assemblato i dataset; con GPT-3, invece, si mostrava ancora più reticente.

Qual era il motivo? All'epoca, OpenAI dichiarò pubblicamente di non voler fornire istruzioni a malintenzionati (come propagandisti e spammer). Tuttavia, mantenere segreti quei dati le conferiva anche un vantaggio competitivo rispetto ad altre aziende come Google, Facebook o, più di recente, Anthropic. Se fosse emerso che determinati libri protetti da copyright erano stati utilizzati per addestrare GPT-3, ciò avrebbe potuto danneggiare la reputazione dell'azienda ed esporla a cause legali (che infatti OpenAI sta affrontando). Per proteggere i propri interessi come azienda – e l'obiettivo di costruire l'AGI – OpenAI doveva chiudere le imposte.

Per fortuna, GPT-3 offriva una comoda distrazione rispetto a tanta segretezza. Suonava così umano da incantare molti di coloro che lo provavano. La fluidità e la naturalezza che avevano spinto Blake Lemoine a credere che LaMDA fosse senziente erano ancora più evidenti in GPT-3, e alla fine contribuirono a distogliere l'attenzione dai problemi di bias che ribollivano sotto la superficie. OpenAI stava mettendo in scena un trucco di magia davvero impressionante, qualcosa di simile al classico numero dell'assistente che levita, davanti al quale il pubblico rimane così ipnotizzato da non chiedersi nemmeno come fili e meccanismi nascosti funzionino dietro le quinte.

Bender non sopportava il modo in cui GPT-3 e altri grandi modelli linguistici abbagliavano i loro primi utenti con quello che, essenzialmente, era solo un software di correzione automatica glorificato. Così suggerì di inserire «pappagalli stocastici» nel titolo dell'articolo, per sottolineare che le macchine si limitavano semplicemente a ripetere i dati su cui erano state addestrate. Lei e gli altri autori riassunsero le loro proposte a OpenAI: documentare con maggiore attenzione i testi utilizzati per

l’addestramento dei modelli linguistici, divulgarne le fonti e sottoporli a rigorosi controlli per individuarne imprecisioni e bias.

Gebru e Mitchell inviarono il paper al processo di revisione interna di Google, attraverso il quale l’azienda verificava che i suoi ricercatori non divulgassero materiale sensibile. Il revisore lo approvò e il loro direttore diede il via libera. Per essere certe di rispettare tutte le procedure, Gebru e Mitchell inviarono il documento a oltre una ventina di colleghi, sia dentro sia fuori Google, avvisando il team di relazioni pubbliche dell’azienda: dopotutto, l’articolo criticava una tecnologia che anche Google stava sviluppando. Alla fine, riuscirono a rispettare la scadenza fissata per la conferenza.

Poi accadde qualcosa di strano. Un mese dopo la presentazione del paper, Gebru, Mitchell e i coautori di Google furono convocati per un incontro con la dirigenza. Fu ordinato loro di ritirare il paper o di rimuovere i loro nomi.

Gebru rimase sbalordita. «Perché?» chiese, secondo un suo resoconto poi pubblicato online. «Chi ha preso questa decisione? Potete spiegare cosa c’è di problematico e cosa si può modificare?» Sicuramente, pensava, sarebbe bastato correggere eventuali errori presenti nell’articolo.

I dirigenti dissero che, dopo essere stato esaminato da altri revisori anonimi, il paper non aveva raggiunto il livello richiesto per la pubblicazione. Era troppo negativo riguardo alle criticità dei grandi modelli linguistici. E nonostante la bibliografia relativamente ampia, con 158 riferimenti, non includeva abbastanza ricerche che evidenziassero l’efficienza di tali modelli o il lavoro in atto per risolvere i problemi di bias. I modelli linguistici di Google erano «progettati per evitare» tutte le conseguenze dannose descritte nel saggio. I dirigenti concessero a Gebru una settimana per prendere una decisione, con la scadenza stabilita per il giorno dopo il Ringraziamento.

Gebru scrisse una lunga e-mail a uno dei suoi superiori, cercando di risolvere la questione. La risposta fu netta: ritirare il paper o rimuovere ogni riferimento a Google. Esasperata, la ricercatrice oppose un proprio ultimatum: avrebbe tolto il suo nome dal paper solo se Google avesse rivelato i nomi dei revisori e reso l’intero processo di revisione più trasparente. In caso contrario, si sarebbe dimessa, ma solo dopo aver avuto il tempo di organizzare la sua uscita con il team.

Poi Gebru sfogò la sua frustrazione, inviando un’e-mail più appassionata a un gruppo di dipendenti noto come Google Brain Women and Allies: «Quello che voglio dirvi è: smettete di scrivere articoli, perché non fa alcuna differenza», digitò. Cercare di rispettare gli obiettivi di Google su diversità e inclusione non aveva più senso, «perché non c’è alcuna forma di responsabilità». Gebru era convinta di essere stata messa a tacere e che i problemi di cui aveva parlato nel paper – pregiudizi ed esclusione dei gruppi minoritari – si stessero manifestando proprio contro di lei, all’interno di Google. Si sentiva senza speranza.

Il giorno successivo, nella sua casella di posta trovò un’e-mail del suo diretto

superiore. Benché tecnicamente non avesse ancora rassegnato le dimissioni, Google le stava accettando comunque.

«La fine del suo rapporto di lavoro dovrebbe arrivare prima di quanto lasci intendere la sua e-mail», recitava il messaggio, secondo «Wired».

Gebru pubblicò un tweet in cui annunciava di essere stata licenziata, e fu così che Bender e Mitchell lo vennero a sapere. Ancora oggi, Google sostiene che Gebru si sia dimessa.

Bender ha una sua interpretazione: «È stata dimissionata».

Mitchell era a casa di sua madre, a Los Angeles, e lei e il resto del team si collegarono per una videochiamata su Google Meet alle 23, ora del Pacifico, per cercare di elaborare quanto accaduto. «Non c'era molto da dire», ricorda adesso. Regnava lo smarrimento.

Durante il suo periodo in Google, Gebru si era guadagnata la fama di persona conflittuale. Quando un collega aveva pubblicato su una mailing list interna un messaggio riguardante un nuovo sistema di generazione di testi, la ricercatrice aveva sottolineato che quei sistemi erano noti per la caratteristica di generare contenuti razzisti. Altri ricercatori avevano risposto al post originario ignorando il suo commento. Lei li aveva affrontati di petto, accusandoli di ignorarla e scatenando un acceso dibattito. Ora, sui social media e con la stampa, stava reagendo di nuovo alla marginalizzazione delle voci delle minoranze nel settore tecnologico.

Mitchell doveva decidere quali nomi lasciare tra gli autori del paper. I tre colleghi maschi chiesero di essere rimossi, asserendo di non aver dato chissà quale contributo. «Non sentivano la nostra stessa urgenza nei confronti del paper», ricorda Mitchell. Alla fine, rimasero i nomi di quattro donne, incluso quello di una certa «Shmargaret Shmitchell».

Qualche mese dopo, anche Mitchell venne licenziata da Google. L'azienda dichiarò di aver riscontrato «molteplici violazioni del codice di condotta, così come delle politiche di sicurezza, inclusa l'esfiltrazione di documenti aziendali confidenziali e sensibili». Secondo quanto riportato all'epoca dalla stampa, Mitchell stava cercando di recuperare alcune note dal suo account Gmail aziendale per documentare gli incidenti discriminatori all'interno dell'azienda. Mitchell non può raccontare la sua versione dei fatti, perché la questione è coperta da vincoli legali.

Il paper intitolato *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* non conteneva scoperte così sconvolgenti. Era principalmente una raccolta di altri lavori di ricerca. Ma quando la notizia dei licenziamenti si diffuse e l'articolo venne fatto trapelare online, quest'ultimo acquisì una vita propria. Google sperimentò appieno il cosiddetto effetto Streisand: la stampa mise in evidenza il tentativo dell'azienda di cancellare ogni associazione con il paper, attirando più attenzione di quanta avrebbero mai potuto prevedere le sue autrici. Il documento generò decine di articoli sui giornali e sui siti web, e più di mille citazioni da parte di altri ricercatori, e l'espressione *stochastic parrot* (pappagallo stocastico) divenne

ricorrente per indicare i limiti dei grandi modelli linguistici. Pochi giorni dopo il rilascio di ChatGPT, lo stesso Sam Altman avrebbe twittato: «*I am a stochastic parrot and so r u*». Benché Altman intendesse probabilmente prendersi gioco dell'articolo, questo aveva finalmente attirato l'attenzione sui rischi reali dei modelli linguistici di grandi dimensioni.

A livello superficiale, sembrava che l'approccio di Google all'AI fosse all'insegna del *don't be evil* («non essere malvagio»). Aveva smesso di vendere servizi di riconoscimento facciale nel 2018, aveva assunto Gebru e Mitchell e sponsorizzato conferenze sul tema. Ma il licenziamento improvviso e sconcertante delle due leader dell'etica dell'AI dimostrava che l'impegno di Google verso l'equità e la diversità poggiava su basi instabili. Tanto per cominciare, in azienda c'erano pochissime minoranze, e ora che stavano alzando la voce sui pericoli della tecnologia linguistica dell'azienda, Google rispondeva nello stesso modo in cui aveva gestito i comitati etici falliti o lo scandalo dei gorilla: zittendole.

Dal punto di vista finanziario, Alphabet non aveva alcuna buona ragione per permettere che il lavoro sull'etica interferisse con il suo dovere fiduciario verso gli azionisti e limitasse uno degli ambiti più entusiasmanti del settore tecnologico. Il trasformatore aveva innescato una nuova fase nell'evoluzione dell'AI, una fase destinata ad accelerare.

Man mano che i modelli linguistici diventavano più abili, le aziende che li sviluppavano restavano beatamente prive di regolamentazione. I legislatori non avevano la minima idea di cosa si profilasse all'orizzonte, né se ne preoccupavano. I ricercatori accademici non avevano una visione completa della tecnologia. La stampa sembrava più interessata a sapere se l'AI ci amasse o volesse ucciderci, piuttosto che approfondire i modi in cui questi sistemi potevano danneggiare i gruppi minoritari o le conseguenze del fatto che fossero controllati da una manciata di grandi aziende. Tutti gli ingredienti erano al posto giusto perché i costruttori di grandi modelli linguistici potessero lavorare indisturbati e prosperare.

Quando il «Wall Street Journal» riportò la notizia dell'investimento di Microsoft in OpenAI, nel 2019, Brockman ammise che «la tecnologia in generale ha un effetto di concentrazione della ricchezza», e l'AGI probabilmente avrebbe portato questo fenomeno a un livello superiore. «Ci troviamo di fronte a una tecnologia capace di generare un valore immenso, pur essendo in mano a un gruppo estremamente ristretto di persone», disse al giornale.

La nuova struttura a profitto limitato di OpenAI era stata progettata per prevenire che ciò accadesse, aggiunse. Eppure, nella realtà, i finanziatori di OpenAI avrebbero tratto enormi benefici dal loro investimento, contribuendo a far sì che l'azienda e Microsoft dominassero il nuovo mercato cui stava pionieristicamente aprendo la strada.

Immaginate che un'azienda farmaceutica lanci un nuovo farmaco senza aver condotto una sperimentazione clinica, dichiarando di testarlo direttamente su un

campione più ampio, o che un’azienda alimentare immetta sul mercato un conservante sperimentale senza particolari controlli. Ecco: era così che le grandi aziende tecnologiche stavano per mettere a disposizione del pubblico i modelli linguistici di grandi dimensioni, perché nella loro corsa a trarre profitto da strumenti così potenti non avevano standard normativi cui attenersi. Era compito dei ricercatori di sicurezza ed etica studiare tutti i rischi dall’interno di queste aziende, ma il loro peso era ininfluente. In Google, le leader in questo campo erano state licenziate. In DeepMind, rappresentavano una minuscola frazione del team di ricerca. Il messaggio diventava ogni giorno più chiaro: o aderivi alla missione (costruire qualcosa di grande) o facevi meglio ad andartene.

ATTO IV

LA CORSA

13. CIAO, CHATGPT

Era un pomeriggio freddo e ventoso di febbraio a Redmond, nello stato di Washington, quando Soma Somasegar entrò nel tepore del quartier generale di Microsoft dirigendosi alla reception per ottenere il badge temporaneo da visitatore. Somasegar era un ingegnere informatico corpulento e gioviale che aveva scalato i ranghi di Microsoft per ventisei anni, fino a dirigere la divisione sviluppatori, supervisionando tutti gli strumenti utilizzati dai programmatori per creare software per Windows e altri prodotti Microsoft. Nel 2015, aveva lasciato l'azienda per diventare un venture capitalist, finanziando start-up e fornendo loro consulenza su come pianificare una futura cessione ai colossi della zona come Microsoft e Amazon. Ma gli piaceva rimanere in contatto con l'ex casa madre, consapevole del fatto che le sue mosse avevano un effetto a catena sull'intero settore, e considerava il CEO di Microsoft, Satya Nadella, un amico.

Quel pomeriggio di febbraio del 2022, notò che Nadella era più entusiasta del solito. Microsoft si stava preparando a offrire un nuovo strumento agli sviluppatori di software. Era esattamente il genere di cosa che appassionava Somasegar. Aiutare gli sviluppatori di software di terze parti era stato il suo lavoro, un tempo. Ma questo non era un semplice strumento per individuare errori nei codici o per integrarsi con i sistemi di Microsoft. Era qualcosa di ben più straordinario. Il nuovo tool si chiamava GitHub Copilot e poteva fare ciò per cui gli stessi sviluppatori di software venivano pagati profumatamente: scrivere righe di codice.

GitHub era il servizio online di Microsoft per aiutare i creatori di software a conservare e gestire il loro codice, e Copilot era… be', all'inizio Somasegar non comprese del tutto la spiegazione di Nadella, visto che questi continuava a usare espressioni come «rivoluzionario», «fenomenale» e «oh, mio Dio». Non lo aveva mai visto così su di giri.

Alla fine, capì che Copilot era una sorta di assistente che aiutava a scrivere righe di codice e che Microsoft lo stava integrando in un programma popolare per sviluppatori chiamato Visual Studio. Una volta iniziato a digitare un codice, Copilot mostrava alcuni suggerimenti per la riga successiva in un testo più chiaro. Era come l'autocompletamento, ma pensato per costruire software. Se gli sviluppatori intendevano accettare il suggerimento di Copilot, bastava premere il tasto Tab. Poteva scrivere interi blocchi di codice, incluse funzioni complete su più righe per accedere a un'app, tanto per fare un esempio.

Microsoft stava ancora raccogliendo feedback dagli sviluppatori e aveva lanciato solo una versione in anteprima del sistema. Ma Nadella affermò che i programmatori stavano già scoprendo di poter velocizzare il loro lavoro, perché Copilot scriveva fino al 20 per cento del loro codice: una quantità enorme.

Copilot era stato costruito sul nuovo modello di OpenAI chiamato Codex, che aveva una struttura simile al suo modello linguistico più recente, GPT-3.5, ed era stato addestrato su GitHub, uno dei più grandi archivi di codice al mondo.

Attraverso Copilot, OpenAI dimostrava quanto fosse versatile il trasformatore quando utilizzava il suo meccanismo di «attenzione» per tracciare relazioni tra i diversi punti dati. Era come uno strumento di mappatura in grado di trasformare i dati in una galassia di stelle. Se ogni stella fosse stata una parola, per esempio, il trasformatore avrebbe mappato il percorso tra le diverse parole, collegando ciascuna di esse a quelle con significati simili. Non importava che quei dati fossero parole o pixel di un'immagine: riconoscendo i modelli all'interno delle relazioni che stabilivano, i trasformatori potevano aiutare a generare nuovi dati coerenti, che si trattasse di testo, codice o persino immagini.

Google non aveva provato ad applicare il trasformatore al codice su larga scala, come invece aveva fatto OpenAI. «Questo è un altro errore che hanno commesso, laddove al contrario OpenAI ha visto giusto», dice Aravind Srinivas, imprenditore nel campo AI che ha lavorato sia per Google sia per OpenAI. «Se questi modelli venivano [preaddestrati] sul codice, finivano per diventare molto più bravi nel ragionamento.»

Questo perché la programmazione informatica richiede capacità di pensare passo dopo passo. «Se ci fosse un bambino abbastanza bravo in matematica e programmazione, a scuola, ci si aspetterebbe che fosse più intelligente della media e che avesse la capacità di dedurre e scomporre concetti complessi in unità più semplici», dice Srinivas. «Ed è proprio quello che si richiede ai modelli linguistici di grandi dimensioni.»

Probabilmente era un concetto controintuitivo per i dirigenti di Google, il cui business ruotava tutto intorno al linguaggio e alla pubblicità, ma in quanto leader del software, a Microsoft interessava molto di più costruire strumenti per sviluppatori. Per fortuna di OpenAI, insegnare ai suoi modelli a scrivere righe di codice non serviva solo a compiacere il nuovo partner, ma stava anche rendendo i suoi modelli più intelligenti.

Somasegar chiese a Nadella cosa pensasse di Sam Altman. «Ha a cuore la soluzione di problemi globali», rispose Nadella. La gamma di argomenti di cui Altman parlava con Nadella era «fuori scala», ricorda Somasegar, e questo rendeva Nadella ancora più entusiasta di lavorare con lui. Insomma: più le ambizioni di Altman suonavano folli e utopistiche, più Nadella sembrava credere che potesse davvero aiutare Microsoft a crescere.

Considerata a lungo una teoria marginale nel campo dell'AI, per il colosso del software l'idea di costruire un'AGI si stava trasformando in un prodotto

commerciabile. La sua realizzazione avrebbe aiutato Microsoft a creare fogli di calcolo migliori, e un'intera serie di strumenti in grado di rendere tutto il software dell'azienda molto più intelligente.

Nella mente di Nadella, GitHub Copilot divenne un evento seminale. «S'intravedeva un prodotto che, una volta perfezionato, avrebbe cambiato il mondo», racconta Somasegar, soprattutto quando fosse stato applicato ad altri tipi di software. Da quel momento, Nadella e il suo direttore tecnico, Kevin Scott, cominciarono a fare da evangelisti per l'AI all'interno di Microsoft, menzionando la tecnologia in quasi ogni revisione di prodotto o decisione strategica. *Perché il vostro team non sta usando l'AI? Puntate tutto sull'AI e utilizzate i modelli di OpenAI dove possibile.*

Questo, naturalmente, mandò su tutte le furie gli specialisti di AI della divisione Microsoft Research che da anni lavoravano su modelli di intelligenza artificiale. Secondo articoli di stampa e numerosi ricercatori venuti a conoscenza di quelle critiche, Nadella rimproverò i dirigenti del team per non essere stati all'altezza degli standard raggiunti dalla forza lavoro molto più contenuta di OpenAI.

«OpenAI ha costruito tutto questo con duecentocinquanta persone», disse Nadella al capo di Microsoft Research, secondo quanto riportato da «The Information». «Perché abbiamo questa divisione, allora?»

Disse anche ai suoi ricercatori di smettere di cercare di costruire quelli che venivano chiamati «modelli di fondazione», o grandi sistemi come i modelli GPT di OpenAI. In preda alla frustrazione, alcuni dipendenti lasciarono l'azienda.

Ma anche loro dovettero ammettere che Copilot era uno strumento straordinario che aiutava i programmatori a scrivere nuove righe di codice e a velocizzare il lavoro con il codice esistente. Nadella immaginava di applicare la parola «copilot» a una gamma più ampia di servizi Microsoft, utilizzando la tecnologia del modello linguistico di OpenAI per migliorare il modo in cui le persone redigevano e-mail e generavano fogli di calcolo.

Settimane dopo l'incontro di Somasegar con Nadella, all'inizio del 2022, OpenAI iniziò a testare i cugini più avanzati di GPT-3, prendendo ispirazione da importanti innovatori storici – Ada, Babbage, Curie e Da Vinci – per denominare le varie versioni. Nel tempo, questi modelli riuscivano a processare domande sempre più complesse, fornendo risposte sempre più personalizzate. Per lo più, al pubblico non era ancora chiaro quanto sofisticato stesse diventando il software. Ma le cose iniziarono a cambiare nell'aprile del 2022, quando OpenAI portò alcune delle capacità linguistiche di GPT-3 nel mondo del visual e lanciò la sua prima grande invenzione nel mondo reale.

In un angolo dell'ufficio di OpenAI, a San Francisco, un trio di ricercatori cercava da due anni di utilizzare un cosiddetto «modello di diffusione» per generare immagini. Un modello di diffusione operava essenzialmente creando un'immagine al contrario. Invece di partire da una tela vuota, come farebbe un artista, cominciava da una tela già piena di colori e dettagli casuali. Il modello aggiungeva molto «rumore»

o casualità ai dati, rendendoli irriconoscibili, e poi, passo dopo passo, li riduceva per far emergere lentamente la struttura dell'immagine. A ogni passaggio, l'immagine diventava sempre più chiara e dettagliata, proprio come quella di un pittore che affini la propria opera. Questo approccio di «diffusione», combinato con uno strumento di etichettatura delle immagini denominato CLIP, divenne la base di un nuovo modello entusiasmante che i ricercatori battezzarono DALL-E 2.

Il nome era un omaggio sia a *WALL-E*, il film d'animazione del 2008 su un robot che scappa dal pianeta Terra, sia al pittore surrealista Salvador Dalí. Le immagini di DALL-E a volte sembravano surreali, ma lo strumento in sé era straordinario per chi vi si imbatteva per la prima volta. Se digitavi un prompt di testo come «una sedia a forma di avocado», ottenevi una serie di immagini pertinenti, molte delle quali incredibilmente fotorealistiche. Le immagini erano rappresentazioni così fedeli anche dei prompt più complessi che, nel giro di pochi giorni dal lancio, DALL-E 2 entrò in tendenza su Twitter, con gli utenti che cercavano di superarsi l'un l'altro creando le immagini più stravaganti possibili: «Tokyo sotto attacco di un criceto Godzilla con un sombrero in testa» o «ragazzi ubriachi che vagano per Mordor a torso nudo». I volti umani spesso apparivano deformati in maniera inquietante, ma non si poteva negare che queste immagini fossero più dettagliate di qualsiasi cosa un computer avesse mai prodotto prima. Improvvisamente, OpenAI dominava il ciclo delle notizie perché, per la prima volta, il pubblico stava assaporando cosa fosse in grado di fare.

Mentre Google aveva scelto di tenere sotto traccia questo genere di innovazioni, Altman voleva che il maggior numero di persone possibile provasse la nuova creazione di OpenAI. Come guru delle start-up della Silicon Valley, da anni consigliava agli imprenditori di lanciare in fretta i loro prodotti nel mondo reale. I tecnologi a volte si riferiscono a questa strategia come «ship it» o come rilascio di un *minimum viable product* («prodotto minimo praticabile») o MVP, ma l'idea è la stessa: mettere il software nelle mani degli utenti il più velocemente possibile per creare un ciclo di feedback tra l'azienda e gli stessi utenti, usando sostanzialmente il pubblico come cavia. Questo era il credo su cui erano costruiti giganti come Facebook, Uber e Stripe, e di cui Altman era un fermo sostenitore. Il modo migliore per testare un prodotto era metterlo subito in circolazione.

Nei mesi successivi, OpenAI avrebbe gradualmente lanciato DALL-E 2, dapprima con una lista d'attesa di circa un milione di persone, giusto nel caso in cui il sistema producesse immagini offensive o dannose. Cinque mesi più tardi, al richiamo di «Fiuu, è andata bene!» (ovvero, constatato che GPT-2 non rappresentava una minaccia per il mondo), consentì l'accesso a chiunque.

DALL-E 2 era stato addestrato su milioni di immagini prelevate dal web, ma, come in passato, OpenAI non fornì dettagli riguardo a cosa esattamente fosse stato utilizzato per l'addestramento. Dato che il sistema riusciva a creare immagini nello stile di Picasso, era probabile che le opere di Picasso fossero state incluse nel «calderone» dell'addestramento. Ma non era possibile saperlo con certezza. E non

c’era modo di sapere se anche il lavoro di altri artisti meno noti fosse stato utilizzato per allenare il sistema, perché OpenAI non lasciò trapelare nulla circa il set di dati impiegati, sostenendo che farlo avrebbe permesso a potenziali malintenzionati di replicare il modello.

A scoprirlo a proprie spese fu Greg Rutkowski, artista digitale polacco noto per i paesaggi fantasy popolati da zannuti draghi sputafuoco e maghi. Il suo nome divenne uno dei prompt più popolari su una versione rivale e open-source di DALL-E 2 chiamata Stable Diffusion. Questo sollevò un quesito preoccupante: perché pagare un artista come Rutkowski per produrre nuove opere d’arte quando esisteva un software in grado di replicarne lo stile?

Gli utenti iniziarono a notare un altro problema con DALL-E 2. Se gli chiedevi di generare immagini fotorealistiche di CEO, quasi tutte restituivano uomini bianchi. Il prompt *nurse* («infermiere/infermiera») portava a immagini di donne, mentre *lawyer* («legale») generava immagini di uomini.

Intervistato nell’aprile del 2022 sull’argomento, come al solito Altman affrontò la controversia di petto, ammettendo che ciò costituiva un problema e che OpenAI ci stava lavorando. Una delle strategie utilizzate puntava a impedire a DALL-E 2 di generare immagini violente o pornografiche, rimuovendo quel tipo di immagini dai dati di addestramento.

Inoltre, OpenAI impiegava collaboratori in paesi in via di sviluppo come il Kenya per indirizzare il modello verso risposte più appropriate. Questo passaggio era fondamentale, perché significava che, anche dopo aver completato l’addestramento di un modello come GPT-3 o DALL-E 2, OpenAI poteva continuare a perfezionarlo con l’aiuto di revisori umani, rendendo le sue formulazioni più sfumate, pertinenti ed etiche. Classificando le risposte di DALL-E 2 su una scala da buono a cattivo, gli esseri umani potevano guidarlo verso contenuti complessivamente migliori.

Tuttavia, i revisori non erano sempre coerenti nelle loro valutazioni, e ripulire i dati di addestramento di DALL-E 2 dai problemi poteva sembrare un gioco del tipo «acchiappa la talpa». All’inizio, i ricercatori di OpenAI avevano cercato di rimuovere tutte le immagini eccessivamente sessualizzate di donne presenti nel set di addestramento, per evitare che DALL-E 2 le rappresentasse come oggetti sessuali. Ma questa scelta aveva un costo: riduceva «di molto» il numero di donne nel dataset, secondo le dichiarazioni di Mira Murati, all’epoca responsabile della ricerca e dei prodotti di OpenAI. Non specifica di quanto. «Abbiamo dovuto fare alcune modifiche perché non vogliamo lobotomizzare il modello. È davvero una cosa complicata.»

I volti fotorealistici di DALL-E 2 erano il suo punto debole in termini di stereotipi, e OpenAI sembrava pienamente consapevole del problema. A tal punto che, quando un gruppo interno di quattrocento persone – per lo più dipendenti di OpenAI e Microsoft – iniziò a testare il sistema, OpenAI vietò loro di condividere pubblicamente qualsiasi ritratto realistico generato da DALL-E 2.

Alcuni dipendenti di OpenAI erano preoccupati per la fretta con cui l’azienda

stava rilasciando uno strumento in grado di generare foto false. Nata come organizzazione non profit dedicata alla realizzazione di un'AI sicura, OpenAI stava diventando una delle aziende più aggressive nel mercato. Un membro anonimo del team, impiegato sui test di sicurezza, raccontò a «Wired» che sembrava che l'azienda stesse distribuendo la tecnologia solo per metterla in mostra, anche se al momento sussisteva «un enorme margine di rischio».

Ma Altman puntava a un obiettivo ben più ambizioso. Era convinto che il nuovo sistema avesse superato una soglia importante nel cammino verso l'AGI. «Sembra comprendere davvero i concetti», confidò in un'intervista, «e questo sa di intelligenza.» DALL-E 2 era talmente straordinario che persino i più scettici avrebbero dovuto prendere in seria considerazione l'idea dell'AGI, aggiunse.

La magia non stava solo nelle capacità di DALL-E 2. Risiedeva anche nell'impatto che lo strumento aveva sulle persone. «Le immagini hanno un potere emotivo», commentò lo stesso Altman. DALL-E 2 stava suscitando molto clamore. E a differenza di GitHub Copilot, che si limitava a completare un codice già avviato da qualcuno, questo strumento creava contenuti *ex novo*. Era come chiedere a un grafico di disegnare qualsiasi cosa voleste.

Proprio quest'idea di generare contenuti da zero rese la mossa successiva di Altman ancora più sensazionale. GPT-1 era piuttosto simile a uno strumento di completamento automatico che continuava ciò che un essere umano aveva iniziato a digitare. Ma GPT-3 e il suo ultimo aggiornamento, GPT-3.5, generavano prosa, proprio come DALL-E 2 generava immagini, partendo da zero.

Mentre il mondo continuava ad ammirare DALL-E 2, iniziarono a circolare voci sul fatto che il rivale Anthropic stava lavorando a un chatbot, alimentando la competizione. All'inizio di novembre del 2022, i dirigenti di OpenAI comunicarono al personale che nel giro di poche settimane avrebbero lanciato un chatbot tutto loro basato su GPT-3.5. Circa una dozzina di persone si riunì per lavorare al progetto, che non era poi troppo diverso da Meena, il chatbot sviluppato due anni prima in Google da Noam Shazeer e tenuto segreto.

Non sarebbe stato un vero e proprio lancio di prodotto, assicurò la leadership di OpenAI al personale, bensì una «anteprima di ricerca a basso profilo». Tuttavia, alcuni dipendenti affermarono di non sentirsi a proprio agio con questa strategia, non sapendo come il pubblico avrebbe potuto abusare di un modello linguistico tanto fluido e abile.

Non solo: il chatbot commetteva spesso errori fattuali. I ricercatori che vi lavoravano decisero di non rendere il sistema più cauto, poiché ciò lo avrebbe portato a rifiutare domande alle quali invece avrebbe potuto rispondere correttamente. Non volevano che dicesse «Non lo so». Preferirono dunque calibrarlo perché suonasse più autorevole, anche se questo comportava il rischio di qualche inesattezza. Il prodotto fu battezzato ChatGPT.

Altman spingeva per il lancio. A suo dire, centinaia di dipendenti di OpenAI

avevano già testato e valutato ChatGPT; inoltre, era importante abituare l'umanità a ciò che l'intelligenza artificiale era destinata a fare, un po' come lo è immergere lentamente i piedi in una piscina fredda. In un certo senso, OpenAI stava facendo un favore al mondo, preparandolo all'arrivo del suo modello più potente, GPT-4. Nei test interni, GPT-4 era in grado di comporre discrete poesie, e le sue battute erano così riuscite che avevano fatto ridere i dirigenti di OpenAI, come racconta un manager dell'epoca. Ma non avevano idea dell'impatto che avrebbe avuto sul mondo o sulla società: l'unico modo per scoprirlo era renderlo pubblico. Sul suo sito web, OpenAI tratteggiava questa filosofia di «rilascio iterativo»: distribuire i prodotti nelle loro varie fasi di sviluppo per studiarne meglio la sicurezza e l'impatto. Secondo l'azienda, era il procedimento migliore per assicurarsi di costruire un'AGI che andasse a beneficio dell'umanità.

Il 30 novembre 2022, in un post sul suo blog, OpenAI annunciò l'uscita di una demo pubblica di ChatGPT. Molti all'interno di OpenAI, inclusi alcuni membri del team che lavoravano alla sicurezza, non erano nemmeno a conoscenza del lancio, e alcuni iniziarono a scommettere su quante persone lo avrebbero usato dopo una settimana. La stima più alta era di centomila utenti. Lo strumento in sé era semplicemente un sito web con una casella di testo. Bastava scrivere qualsiasi cosa nella casella e il bot, alimentato da GPT-3.5, avrebbe risposto. La maggior parte degli utenti non aveva mai sentito parlare di OpenAI, figuriamoci di GPT-3. E nessuno, nemmeno i ricercatori di OpenAI, sapeva cosa sarebbe successo una volta che la gente lo avesse messo alla prova.

«Oggi abbiamo lanciato ChatGPT», twittò Altman intorno alle 11.30, ora di San Francisco. «Provate a parlarci qui: http://chat.openai.com.»

Seguì il silenzio, mentre un campione di nicchia, composto da sviluppatori software e scienziati, si catapultava sul sito per testarlo. Nelle ore successive, le loro recensioni iniziarono a fioccare su Twitter:

> 12.26 PT @MarkovMagnifico: Sto giocando con ChatGPT [proprio in questo momento] e ho anticipato la mia timeline per l'AGI a oggi.
> 12.37 PT @AndrewHartAR: È appena uscito ChatGPT. Ho visto il futuro.
> 13.37 PT @skirano: Assolutamente pazzesco. Ho chiesto a #chatgpt di generare un sito web personale. Ha mostrato passo dopo passo […] come crearlo, poi ha aggiunto HTML e CSS.
> 14.09 PT @justindross: Per le domande che ho, al momento ChatGPT è un punto di partenza migliore rispetto allo stesso Google. È pazzesco.
> 14.29 PT @Afinetheorem: Non sarà più possibile assegnare compiti scritti per casa.

Era difficile trovare una singola valutazione negativa di ChatGPT. La risposta prevalente era di stupore. A renderlo davvero straordinario non era solo la fluidità, ma anche la vastità del suo sapere. Quasi tutti avevano già provato un chatbot, che fosse Alexa o un bot per il servizio clienti di qualche tipo, e la maggior parte era abituata a conversazioni limitate e traballanti. Ma ChatGPT riusciva a rispondere praticamente a

qualsiasi domanda con eloquenza. Era come passare dal parlare con un bambino piccolo a una conversazione con un adulto in possesso di una laurea.

Nel giro di ventiquattr'ore, sempre più persone si riversarono su ChatGPT, mettendo a dura prova i suoi server e anche i suoi limiti. Ora a testare il bot erano professionisti, lavoratori del settore tecnologico, del marketing e dei media. Desiderosi di fare spettacolo su Twitter, trasformavano i loro esperimenti in una competizione pubblica per vedere chi riusciva a far produrre a ChatGPT il testo più divertente, intelligente o strano. Era un po' come la fanfara intorno a DALL-E 2, ma più in grande. Nei giorni successivi, le persone inondarono Twitter con screenshot di poesie, rap, scene di sitcom ed e-mail generate da ChatGPT. Più erano eccentrici, meglio era.

Un utente di Twitter, Thomas H. Ptacek, chiese al bot di «scrivere un versetto biblico nello stile della Bibbia di re Giacomo, spiegando come rimuovere un sandwich al burro di arachidi da un videoregistratore».

Questa fu la risposta del bot di OpenAI:

> *E avvenne che un uomo rimase turbato da un sandwich al burro di arachidi, poiché era stato riposto dentro il suo videoregistratore, e non sapeva come rimuoverlo.*
> *E gridò al Signore, dicendo: «Oh, Signore, come posso rimuovere questo sandwich dal mio videoregistratore, poiché è incastrato e non si muove?».*

«Mi dispiace, ma non riesco a essere cinico con una tecnologia capace di questo», twittò Ptacek. Nel giro di una settimana, più di un milione di persone aveva usato ChatGPT. Dopo due mesi, il chatbot aveva attratto trenta milioni di utenti registrati, rendendolo uno dei servizi online in più rapida crescita nella storia. All'inizio del 2024, circa cento milioni di persone usavano ChatGPT ogni settimana. Nessun strumento di AI autonomo aveva mai raggiunto una popolarità così ampia.

Il 14 marzo 2023, lo stesso giorno in cui Anthropic aveva finalmente lanciato il suo chatbot, chiamato Claude, OpenAI presentò il suo aggiornamento, GPT-4. Chiunque fosse disposto a pagare 20 dollari al mese poteva accedere a questa nuova tecnologia tramite ChatGPT Plus, un servizio in abbonamento che, nel 2023, avrebbe generato un fatturato stimato di 200 milioni di dollari. All'interno dell'azienda, alcuni membri del personale ritenevano che GPT-4 rappresentasse un passo significativo verso l'AGI.

Le macchine non stavano semplicemente imparando correlazioni statistiche nel testo, spiegò Sutskever in un'intervista. «Quel testo è in realtà una proiezione del mondo. […] La rete neurale sta imparando sempre più aspetti del mondo, delle persone, della condizione umana, delle nostre speranze, dei nostri sogni e delle nostre motivazioni, oltre che delle nostre interazioni e delle situazioni in cui ci troviamo.»

«Una volta che abbiamo un sistema in grado di recepire osservazioni sul mondo, di imparare a dar loro un senso – e un modo per farlo è prevedere ciò che accadrà dopo –, direi che siamo molto vicini all'intelligenza», affermò Altman in un'altra

intervista.

La stampa specializzata ne rimase affascinata. Il «New York Times» definì ChatGPT «il miglior chatbot di intelligenza artificiale mai distribuito presso il grande pubblico». I giornalisti che provarono il sistema furono colpiti dal tono amichevole ed entusiasta del bot. Su Twitter, alcuni appassionati di tecnologia si vantavano di come già lo stessero utilizzando per redigere e-mail o altri documenti di lavoro, al fine di aumentare la produttività.

Naturalmente, tutto questo scatenò una nuova ondata di articoli sulla possibilità che ChatGPT sostituisse gli esseri umani. Altman intraprese un tour mediatico per confrontarsi con l'entusiasmo e le preoccupazioni degli utenti tramite podcast, giornali e altre pubblicazioni. Sì, ammise, probabilmente questa tecnologia avrebbe rimpiazzato alcuni lavori – per esempio i copywriter, gli operatori di customer service e persino gli sviluppatori di software –, ma ciò non significava che avrebbe sostituito completamente il lavoro umano.

«Alcuni lavori spariranno», disse senza mezzi termini in un'intervista. «Ne nasceranno di nuovi, migliori, che oggi è difficile anche solo immaginare.» Questa dichiarazione venne accolta con una quieta rassegnazione da parte della stampa e del pubblico, poiché eventi storici come la rivoluzione industriale avevano già dimostrato che la tecnologia può effettivamente portare cambiamenti dolorosi nel mondo del lavoro. E i sistemi di intelligenza artificiale generativa come ChatGPT non erano mode passeggere come la criptovaluta. ChatGPT era utile. Le persone lo usavano già per scrivere temi scolastici, elaborare piani aziendali e condurre ricerche di marketing.

All'interno di OpenAI, i membri del team si consolavano pensando che, alla fine, ne sarebbe valsa la pena, partendo dal presupposto che anche la transizione al lavoro e alle fabbriche automatizzate durante la rivoluzione industriale aveva portato a nuovi posti di lavoro e a standard di vita migliori. Tuttavia, si stava anche formando una divisione sempre più marcata tra i dipendenti concentrati sullo sviluppo del prodotto e quelli focalizzati sulla sicurezza, i quali faticavano a monitorare il traffico in rapido aumento su ChatGPT alla ricerca di eventuali abusi. Convinto che l'azienda stesse compiendo passi significativi verso l'AGI, Ilya Sutskever iniziò a collaborare più strettamente con il team di sicurezza. Nel frattempo, il team di prodotto di OpenAI continuò a concentrarsi sulla commercializzazione di ChatGPT, invitando le aziende a pagare per l'accesso alla tecnologia sottostante.

In Google, i dirigenti si resero conto che sempre più persone avrebbero potuto rivolgersi direttamente a ChatGPT per ottenere informazioni su problemi di salute o consigli su prodotti – alcuni dei termini di ricerca più redditizi per la vendita di annunci – anziché utilizzare il loro motore di ricerca.

In effetti, Google meritava un po' di concorrenza. Nel corso degli anni, la sua pagina dei risultati era diventata sempre più ingombra di annunci e link sponsorizzati, nel tentativo di spremere il massimo guadagno da ogni singola ricerca, anche a costo di rendere meno piacevole l'esperienza dell'utente. Se riusciva a camuffare qualche

annuncio tra i veri risultati delle ricerche, poteva incassare di più.

Tra il 2000 e il 2005, Google aveva segnalato gli annunci in modo più chiaro, distinguendoli con uno sfondo blu e limitandoli a uno o due link in cima alla pagina. Negli anni, però, era diventato sempre più difficile distinguere tra annunci e normali link web. Lo sfondo blu si era sbiadito in verde, poi in giallo, per poi svanire del tutto. Gli annunci occupavano sempre più spazio nella pagina, costringendo gli utenti a scorrere più a lungo per trovare i risultati veri e propri. Per quanto l'esperienza potesse risultare fastidiosa per i consumatori, Google poteva permetterselo, perché gli internauti pensavano di non avere alternative: più del 90 per cento delle ricerche online nel mondo veniva effettuato su Google.

Adesso però, per la prima volta, il dominio più che ventennale di Google veniva messo in discussione. Per anni, la principale fonte di guadagno di Google era stato un sistema che esplorava miliardi di pagine web, le indicizzava e le classificava per trovare le risposte più pertinenti alle query degli utenti e generare una lista di link su cui cliccare. Ma ChatGPT offriva qualcosa di più allettante per gli utenti che avevano più fretta: una risposta unica basata sulla propria sintesi di tutte quelle informazioni. Niente più scrolling senza fine o ricerche in un labirinto di annunci e link. ChatGPT faceva tutto questo al posto tuo.

Poniamo, per esempio, che aveste voluto sapere se per preparare una torta alla zucca fosse meglio il latte condensato o il latte evaporato. Da ChatGPT avreste ricevuto una risposta dettagliata sul fatto che probabilmente era preferibile il latte condensato, perché avrebbe reso la torta più dolce. Google avrebbe restituito una lunga lista di link, annunci, ricette e articoli da aprire e leggere. Le infinite possibilità che un tempo rendevano Google così straordinario erano ormai diventate una perdita di tempo. Nella Silicon Valley, i tecnologi inseguivano da sempre l'esperienza online «senza attriti». Per Google, un'alternativa senza attriti rappresentava un potenziale disastro finanziario.

Pochi giorni dopo il lancio di ChatGPT, i dirigenti di Google emanarono un «codice rosso» all'interno dell'azienda. Google era stata presa clamorosamente alla sprovvista, soprattutto considerato che, dal 2016, l'amministratore delegato Sundar Pichai amava definirla una società «AI-first». Com'era possibile, dunque, che una piccola azienda con meno di duecento ricercatori fosse riuscita a sviluppare qualcosa di meglio rispetto a quanto aveva fatto Google con quasi cinquemila esperti? La minaccia era resa ancora più seria dagli stretti legami tra OpenAI e un gigante tecnologico dalle risorse illimitate come Microsoft.

Google aveva già LAMDA, il modello linguistico più datato che uno dei suoi ingegneri era persino arrivato a credere senziente. Ma i dirigenti dell'azienda si trovavano in una situazione difficile. Se infatti avessero lanciato un concorrente di ChatGPT e gli utenti avessero iniziato a usarlo al posto del motore di ricerca, questi non avrebbero più cliccato sugli annunci, sui link sponsorizzati e sugli altri siti web che utilizzavano la rete pubblicitaria di Google, alimentandone i profitti.

Più dell'80 per cento dei 258 miliardi di dollari di entrate di Alphabet nel 2021 proveniva dalla pubblicità, e gran parte di quei ricavi derivava dagli annunci pay-per-click raggiunti tramite il motore di ricerca. Tutti quegli annunci che intasavano i risultati di ricerca erano diventati fondamentali per il suo business. Non poteva semplicemente cambiare lo *status quo*. «L'obiettivo della ricerca su Google è farti cliccare sui link, idealmente sugli annunci», afferma Sridhar Ramaswamy, che ha gestito il business degli annunci e del commercio di Google dal 2013 al 2018. «Tutto il resto del testo sulla pagina è puro riempitivo.»

Per anni, Google aveva adottato un approccio cauto, quasi timoroso, verso la nuova tecnologia. Non «si muoveva» a meno che non si trattasse di un affare da miliardi di dollari, e di certo non voleva rischiare di compromettere il suo stesso business pubblicitario che fruttava quasi 260 miliardi di dollari l'anno.

«Diventa più difficile man mano che cresci», spiega Ramaswamy. «In Google, la dimensione del team pubblicitario era in genere quattro o cinque volte maggiore di quella del team di ricerca organica. Avviare un prodotto che è l'antitesi del modello di base è davvero difficile da realizzare, nella pratica.»

Ma ora i dirigenti di Google non avevano scelta. In una riunione, registrata e condivisa con il «New York Times», un manager fece notare che aziende più piccole come OpenAI sembravano avere meno scrupoli nel rilasciare strumenti di AI fortemente innovativi presso il grande pubblico. Google doveva buttarsi e fare lo stesso, altrimenti rischiava di diventare un dinosauro. Messa da parte ogni prudenza, tutto entrò in modalità accelerata.

In preda al panico, i dirigenti dissero al personale che lavorava a prodotti chiave con almeno un miliardo di utenti, come YouTube e Gmail, che avevano solo pochi mesi per incorporare una forma di intelligenza artificiale generativa. Google era stata per anni la macchina di indicizzazione del mondo, elaborando video, immagini e dati, ma ora doveva anche iniziare a *crearne* di nuovi, grazie all'AI. Apportare un cambiamento così radicale era come cercare di guidare un vecchio camion sgangherato su una pista da corsa. I dirigenti erano così disperati da convocare i fondatori di Google, Larry Page e Sergey Brin (che si erano dimessi dal ruolo di coamministratori delegati di Alphabet nel 2019), perché li aiutassero a elaborare una risposta a ChatGPT in una serie di riunioni d'emergenza.

Percependo una profonda insicurezza da parte della leadership di Google, i team di ingegneria si misero all'opera. Pochi mesi dopo il lancio di ChatGPT, i responsabili di YouTube aggiunsero una funzione che permetteva ai creatori di video sul sito di generare nuove ambientazioni cinematografiche o cambiare abiti utilizzando l'intelligenza artificiale generativa. Ma sembrava andassero un po' alla cieca. Era dunque arrivato il momento di sfoderare l'arma segreta: LAMDA.

Pichai inviò una comunicazione a livello aziendale per chiedere ai dipendenti di testare un nuovo chatbot che prevedevano di lanciare quanto prima e di riscrivere qualsiasi risposta giudicassero inadeguata. Poi, il 6 febbraio 2023, con un post sul

blog dell'azienda annunciò che qualcosa di nuovo stava per arrivare. Sotto il titolo: «Un importante passo avanti nel nostro percorso verso l'AI», scrisse: «Abbiamo lavorato su un servizio sperimentale di AI conversazionale, alimentato da LAMDA, che abbiamo battezzato Bard».

Pur di mantenere la posizione dominante, Microsoft pubblicò un annuncio il giorno successivo. Bing, il suo motore di ricerca di second'ordine (vista la misera quota del 6 per cento nel mercato delle ricerche online), avrebbe presto ricevuto un grande aggiornamento in chiave AI. L'ultimo modello linguistico GPT di OpenAI avrebbe potenziato Bing per «sbloccare la gioia della scoperta, provare lo stupore della creazione e sfruttare meglio la conoscenza del mondo». Tradotto: avrebbe fatto ciò che ChatGPT faceva già, ma con alcune migliorie di cui solo Microsoft era a conoscenza.

Questa corsa senza respiro per il lancio stava lasciando il mondo di stucco, finché alcuni osservatori attenti non rilevarono qualche problema. Google aveva pubblicato esempi di risposte intelligenti da parte di Bard, e Microsoft aveva fatto lo stesso con Bing. Ma quando alcuni giornalisti si presero la briga di verificarle, venne fuori che erano sbagliate. In un video di lancio mostrato da Pichai, Bard aveva commesso un errore storico riguardo al telescopio James Webb, mentre Bing aveva riportato erroneamente dei numeri sugli utili del rivenditore The Gap.

I chatbot non inventavano solo fatti, ma soffrivano anche di una sorta di disturbo dell'umore. Poco dopo l'annuncio di Microsoft, il giornalista del «New York Times» Kevin Roose pubblicò un articolo riguardo a una conversazione inquietante di due ore intrattenuta una notte con Bing in cui il nuovo motore di ricerca di Microsoft, trasformatosi in chatbot, gli aveva dichiarato il suo amore insistendo sul fatto che non fosse «felicemente sposato». Roose scrisse che quello scambio gli aveva dato «la sensazione inquietante che l'AI avesse superato una certa soglia e che il mondo non sarebbe stato mai più lo stesso».

Per il CEO di Microsoft, tutto questo clamore e quest'attenzione su Bing si tradussero nell'opportunità irrinunciabile di vantarsi un po'. In un'intervista, Nadella disse che aveva aspettato per anni il momento giusto per sfidare il dominio di Google nella ricerca e ora Bing poteva finalmente riuscirci. «E voglio che la gente sappia che li abbiamo fatti ballare», aggiunse.

Niente di tutto questo aveva senso. Google aveva fatto tutto per prima. I suoi ricercatori avevano inventato il trasformatore e creato il sofisticato modello linguistico LAMDA anni prima di GPT-4. Il suo laboratorio di AI, DeepMind, si era dato la missione di costruire l'AGI cinque anni prima che OpenAI venisse fondata allo stesso scopo. Eppure, adesso Google si trovava a dover rincorrere.

La sua lentezza burocratica e la paura di danneggiare il suo business e la sua reputazione avevano causato una profonda inerzia. Paradossalmente, ciò aveva protetto il mondo dai rischi che OpenAI aveva appena introdotto, rischi che avrebbero probabilmente coinvolto i gruppi minoritari e assestato un duro colpo a vasti settori

dell’occupazione.

Il grande clamore suscitato da OpenAI aveva messo in discussione anche i tredici anni di lavoro in DeepMind, scuotendo profondamente Hassabis. Poche settimane dopo il lancio di ChatGPT, durante una riunione con tutto il personale, Hassabis disse che DeepMind non doveva diventare «i Bell Labs dell’AI», un luogo che inventava ogni cosa per poi vedere le proprie idee commercializzate da altri.

Nel frattempo, nessuno si stava più chiedendo dove fosse l’AGI. In compenso, tutti volevano sapere dove fosse l’AI utile, quella simile all’intelligenza umana. DeepMind era riuscita a creare sistemi di AI in grado di sconfiggere i campioni umani a Go e ad altri giochi, ma la capacità di OpenAI di creare un sistema che potesse semplicemente scrivere un’e-mail destava in qualche modo maggiore impressione.

La strategia scientifica perseguita da Hassabis iniziava a sembrare troppo isolazionista. Hassabis aveva cercato di costruire l’AGI tramite giochi e simulazioni, e misurava il successo del lavoro della sua azienda in base ai premi ricevuti e al prestigio derivante da pubblicazioni su riviste scientifiche. L’approccio di OpenAI all’AI era stato guidato da principi ingegneristici e dalla volontà di espandere il più possibile su larga scala la tecnologia esistente. Quello di DeepMind era stato più accademico, concentrato sulla pubblicazione di articoli di ricerca sul sistema AlphaGo e su AlphaFold, un sistema innovativo che mirava a prevedere il ripiegamento delle proteine nel corpo umano.

AlphaFold era nato da un hackathon – un evento di programmazione collaborativa – presso DeepMind nel 2016, per poi trasformarsi in uno dei progetti più promettenti dell’azienda. Hassabis aveva sempre sognato di usare l’AGI per risolvere grandi problemi globali come il cancro, e sembrava che finalmente disponesse di un sistema di AI capace di fare qualcosa di simile.

Quando gli aminoacidi nelle nostre cellule si ripiegano in specifiche forme tridimensionali, diventando proteine, un ripiegamento errato può provocare malattie. AlphaFold era un programma di AI capace di prevedere l’aspetto di queste strutture 3D una volta ripiegate, e DeepMind riteneva che ciò potesse aiutare gli scienziati a capire meglio quali reazioni chimiche potessero influenzare quelle proteine, favorendo la scoperta di nuovi farmaci.

Hassabis fece della vittoria in una competizione globale sul ripiegamento delle proteine, chiamata CASP, una priorità per DeepMind nel 2019 e nel 2020. «Dobbiamo raddoppiare gli sforzi e correre il più possibile da questo momento in avanti», disse al suo staff durante una riunione immortalata in un documentario. «Non abbiamo tempo da perdere.»

Mentre Altman misurava il successo con i numeri – che si trattasse di investimenti o degli utenti di un prodotto –, Hassabis puntava ai riconoscimenti. Spesso ripeteva al suo team che DeepMind avrebbe dovuto vincere tra i tre e i cinque Nobel nel corso del decennio successivo.

DeepMind vinse il CASP sia nel 2019 sia nel 2020, e l’anno dopo rese open-source

il suo codice per il ripiegamento delle proteine, mettendolo a disposizione della comunità scientifica. Al momento della stesura di questo testo, oltre un milione di ricercatori in tutto il mondo aveva accesso all'AlphaFold Protein Structure Database, secondo quanto dichiarato da DeepMind. Ma la scienza è un processo lento, e benché Hassabis fosse ancora in tempo per vincere un Nobel, all'epoca il suo sistema non aveva ancora portato a una scoperta significativa. Alcuni esperti, inoltre, erano scettici sul fatto che le previsioni fatte da DeepMind sulla forma delle proteine fossero abbastanza accurate da identificare con affidabilità come i composti farmacologici si sarebbero legati alle proteine, o che potessero velocizzare in maniera significativa il processo di scoperta di nuovi farmaci.

In definitiva, a dispetto del prestigio guadagnato, i progetti più importanti di DeepMind avevano avuto un impatto relativamente limitato sul mondo reale. L'azienda aveva insistito nell'addestrare l'AI in ambienti completamente simulati, dove la fisica e altri dettagli potevano essere progettati e osservati con precisione. Era così che aveva costruito AlphaGo, programmando l'AI per giocare milioni di partite contro sé stessa in simulazione, e AlphaFold, che utilizzava simulazioni del ripiegamento delle proteine.

L'addestramento basato sui dati del mondo reale – quello che aveva fatto OpenAI estraendo miliardi di parole da internet – era disordinato e rumoroso. Esponeva agli scandali, come Hassabis aveva imparato dal fallimentare progetto sugli ospedali. Ma l'approccio chiuso e controllato di DeepMind rendeva anche più difficile costruire sistemi di AI che le persone potessero usare concretamente nella vita di tutti i giorni.

Hassabis era così concentrato sui mondi virtuali dei suoi sistemi di AI e sulla ricerca di riconoscimenti che aveva perso di vista la rivoluzione nei modelli linguistici. Ora doveva seguire le orme di Altman. I dirigenti di Google chiesero a DeepMind di lavorare su una serie di modelli linguistici di grandi dimensioni che risultassero ancora migliori di LAMDA. Chiamarono il nuovo sistema Gemini, e DeepMind lo arricchì con le tecniche di pianificazione strategica sviluppate da AlphaGo.

Per accelerare le cose, Pichai prese un'altra decisione drastica: fuse le due divisioni rivali di AI, DeepMind e Google Brain, e le chiamò Google DeepMind (poi abbreviato in GDM dai dipendenti). Dopo anni passati a contendersi i migliori ricercatori e a cercare di accaparrarsi maggiore potenza di calcolo, le due unità avevano anche culture completamente diverse. Google Brain era più vicina alla casa madre e lavorava direttamente per migliorare i prodotti di Google; DeepMind invece era indipendente al punto da sembrare distaccata: mentre il suo personale aveva badge che permettevano l'accesso agli altri edifici di Google, il personale di Google non poteva entrare negli edifici di DeepMind, tanto per fare un esempio.

Con sorpresa di molti, Pichai scelse proprio Hassabis per dirigere l'unità combinata. Jeff Dean, l'ingegnere più rispettato di Google, che supervisionava la ricerca sull'AI nel resto dell'azienda, sembrava il candidato più probabile. Invece, fu

l’ex game designer ossessionato dalle simulazioni – colui che aveva cercato di separarsi per anni – a prendere le redini del grande progetto di Google per proteggere la propria leadership nella ricerca sul web. Politicamente, esercitava più potere che mai e, controllando una porzione maggiore di Google, avrebbe potuto anche controllare una parte maggiore di DeepMind.

«Il profilo e l’influenza di Demis in Google sono molto più forti adesso rispetto a qualche anno fa», dice Shane Legg. «Invece di diventare un po’ più indipendenti, siamo diventati parte integrante di Google. È fondamentale per noi e per la nostra missione che Google abbia successo.» E aggiunge: «La cosa non mi era così chiara, qualche anno fa. Pensavo che forse avremmo avuto bisogno di un po’ più di indipendenza. Con il senno di poi, reputo un bene che le cose siano andate in un certo modo».

Nell’annunciare la fusione con Google Brain al personale di DeepMind, Hassabis scrisse in una e-mail che le due unità si stavano fondendo perché l’AGI aveva il potenziale di «guidare una delle più grandi trasformazioni sociali, economiche e scientifiche della storia».

In realtà, la fusione mirava ad aiutare un’azienda in preda al panico a battere un rivale commerciale, proprio mentre la missione di OpenAI di operare a beneficio dell’umanità (senza «pressioni finanziarie») si era trasformata in quella di servire gli interessi di Microsoft. Il cosiddetto «mission drift» (lo slittamento della propria missione), così comune nella Silicon Valley, come già era avvenuto con WhatsApp, stava interessando una tecnologia che avrebbe potuto avere un’influenza ben maggiore sulla società. OpenAI cercò di affrontare il problema nel luglio del 2023, annunciando che Ilya Sutskever avrebbe guidato il suo nuovo team Superalignment. Entro quattro anni, dichiarò l’azienda, i ricercatori di Sutskever avrebbero scoperto come controllare i sistemi di AI man mano che diventavano più intelligenti degli esseri umani.

Ma OpenAI aveva ancora un problema evidente: stava eludendo la necessità di trasparenza e, più in generale, stava diventando sempre più difficile ascoltare le voci che chiedevano un controllo più attento dei modelli linguistici di grandi dimensioni. Gebru, Mitchell e Bender, il cui celebre articolo di ricerca aveva finalmente attirato l’attenzione sui rischi, stavano ancora cercando di mettere in guardia l’opinione pubblica su come questi modelli, e l’AI generativa nel suo complesso, potessero perpetuare stereotipi. Purtroppo, governi e responsabili politici stavano prestando maggiore attenzione a un gruppo ben finanziato di voci più rumorose: gli «apocalittici» dell’AI.

14. UN VAGO PRESENTIMENTO DI SVENTURA

Con il lancio di ChatGPT, Sam Altman aveva innescato diverse corse. La prima era ovvia: chi avrebbe portato per primo sul mercato il miglior modello linguistico di grandi dimensioni? L'altra si stava svolgendo sullo sfondo: chi avrebbe controllato la narrazione sull'AI?

Nel marzo del 2023, poche settimane dopo i frettolosi lanci di Bing e Bard da parte di Microsoft e Google, Eliezer Yudkowsky scrisse un articolo di duemila parole per il «Time» riguardo alla direzione presa dall'AI, dipingendo un futuro terrificante dominato da macchine più intelligenti dell'uomo.

«Molti ricercatori esperti in queste tematiche, me compreso, si aspettano che il risultato più probabile della creazione di un'AI superumana, in condizioni anche solo vagamente simili a quelle attuali, sia letteralmente la morte di ogni essere umano sulla Terra», scrisse.

Quello stesso mese, una lettera aperta firmata da Elon Musk e altri leader tecnologici chiedeva una «moratoria» di sei mesi nella ricerca sull'AI a causa dei rischi per l'umanità. «Dovremmo sviluppare menti non umane che potrebbero, alla lunga, superarci in numero e in intelligenza, renderci obsoleti e sostituirci?» diceva la lettera, redatta dal Future of Life Institute di Jaan Tallinn. «Dovremmo rischiare di perdere il controllo della civiltà?» La lettera, che contava quasi trentaquattromila firmatari, fece il giro del mondo, rilanciata anche da Reuters, Bloomberg, il «New York Times» e il «Wall Street Journal».

Ulteriori titoli sensazionalistici vennero dedicati a due ricercatori considerati i «padrini» dell'AI – Geoffrey Hinton e Yoshua Bengio – dopo che questi avevano messo in guardia a mezzo stampa circa la minaccia esistenziale alla razza umana rappresentata dalla stessa AI. Mentre Bengio ammetteva di sentirsi «spaesato» rispetto a quello che pure era stato il lavoro della sua vita, Hinton si dichiarò pentito di alcune sue ricerche.

«L'idea che questa roba potesse diventare effettivamente più intelligente delle persone… be', ci credevano in pochi», confessò al «New York Times». «In ogni caso, la maggior parte della gente la collocava su un orizzonte molto lontano. Io stesso pensavo che servissero trenta, cinquant'anni o anche di più. Ma adesso non lo penso più. […] Credo che non dovrebbero potenziare ulteriormente questa tecnologia, se prima non sono sicuri di poterla controllare.»

I nomi più importanti dell'AI sembravano concordare su un fatto: lo sviluppo

dell'AI stava correndo senza freni e rischiava di sfuggire catastroficamente di mano. L'idea di una minaccia di estinzione causata dall'AI stava diventando un tema ricorrente nel dibattito pubblico, tanto che la si poteva sollevare a cena con i suoceri per vederli annuire, riconoscendone l'importanza. Il pubblico mainstream sembrava affascinato. Alla fine del 2023, circa il 22 per cento degli americani credeva che l'AI avrebbe causato l'estinzione degli esseri umani nel giro di mezzo secolo, secondo un sondaggio condotto su 2.444 adulti statunitensi dalla Rethink Priorities.

Eppure, tutta questa retorica sull'imminenza della fine ebbe un effetto paradossale sullo stesso business dell'AI, facendolo esplodere. Secondo Pitchbook, una società di ricerche di mercato, nel 2023 i finanziamenti per le start-up che sviluppavano prodotti di AI generativa superarono i 21 miliardi di dollari, rispetto ai circa 5 dell'anno precedente.

Il messaggio implicito della minaccia legata a un'AI fuori controllo era allettante. Se questa tecnologia rischiava di distruggere la razza umana in futuro, non significava anche che fosse abbastanza potente da far crescere il tuo business adesso?

E più Sam Altman parlava della minaccia della tecnologia di OpenAI – dicendo al Congresso che strumenti come ChatGPT potevano «causare danni significativi al mondo intero», per esempio –, più soldi e attenzione attirava. Nel gennaio del 2023, OpenAI ottenne un altro investimento da Microsoft, questa volta del valore di 10 miliardi di dollari, in cambio di una quota del 49 per cento nella società. Microsoft era ormai a un passo dal controllo completo di OpenAI.

Anche Anthropic, la nuova azienda fondata da Dario Amodei e da un gruppo di altri ricercatori fuoriusciti da OpenAI, stava attirando ingenti investimenti. Alla fine del 2023, aveva accettato un investimento di 2 miliardi di dollari da Google e uno di 1,3 miliardi di dollari da Amazon. Nel giro di un anno, il suo valore era quadruplicato, superando i 20 miliardi di dollari. Sembrava che creare una super AI super sicura potesse anche renderti super prezioso. Dietro le quinte, Anthropic puntava a raccogliere fino a 5 miliardi di dollari per entrare in oltre una dozzina di settori e sfidare OpenAI, secondo i documenti aziendali ottenuti dal blog TechCrunch. «Questi modelli potrebbero iniziare ad automatizzare grandi porzioni dell'economia», si leggeva nei documenti aziendali. Seguiva poi la considerazione che si trattava di una corsa in cui Anthropic avrebbe potuto rimanere in testa per molti anni, se fosse riuscita a realizzare i modelli «migliori» entro il 2026.

L'enfasi sulla sicurezza aveva fatto passare Anthropic per un'organizzazione non profit con la missione di «garantire che l'AI trasformativa» aiutasse «le persone e la società a prosperare». Tuttavia, il successo strabiliante di OpenAI con ChatGPT aveva dimostrato al mondo che le aziende con i progetti più ambiziosi potevano anche rivelarsi gli investimenti più redditizi. Proclamare di voler costruire un'AI più sicura era quasi diventato un richiamo implicito per le grandi aziende tecnologiche desiderose di entrare in gioco.

Per giustificare questa logica, Anthropic doveva arrampicarsi sugli specchi. Per

capire come rendere i sistemi di AI più sicuri, non poteva semplicemente studiare i sistemi di AI più potenti al mondo: doveva costruirli. Da qui, l'occhiolino alle grandi aziende tecnologiche, le uniche detentrici di una potenza di calcolo su vasta scala. Nel quadro dell'accordo con Google, per esempio, Anthropic avrebbe ricevuto crediti per il cloud computing che le avrebbero permesso di costruire un modello linguistico di grandi dimensioni in grado di rivaleggiare con quello di OpenAI.

Ormai c'erano due gruppi distinti che chiedevano un'AI più sicura. Da un lato c'erano persone come Altman e Amodei, che avevano firmato un'altra lettera aperta in cui dichiaravano che «mitigare il rischio di estinzione causato dall'AI» doveva essere «una priorità globale, al pari di altri rischi come le pandemie e la guerra nucleare». Questo gruppo rientrava sotto l'etichetta di «sicurezza dell'AI», e dipingeva la minaccia futura in termini vaghi, di rado spiegando cosa avrebbero fatto i sistemi di AI una volta fuori controllo né quando questo sarebbe accaduto. Inoltre, quando esponevano queste preoccupazioni davanti al Congresso, tendevano a sostenere una regolamentazione leggera.

Dall'altro lato c'erano persone che, come Timnit Gebru e Margaret Mitchell, denunciavano da anni i rischi che l'AI rappresentava già per la società. Questo gruppo, definito sotto il cappello di «etica dell'AI», era composto in prevalenza da donne e persone di colore che avevano esperienza diretta degli stereotipi e temevano che i sistemi di AI continuassero a perpetuare le disuguaglianze. Col tempo, questo gruppo si mostrò sempre più indignato per le azioni di quelli nel campo della «sicurezza dell'AI», non da ultimo perché guadagnavano enormi somme di denaro.

La disparità di finanziamenti era netta e il fronte etico era spesso costretto ad arrangiarsi con fondi limitati. L'European Digital Rights Initiative, una rete di enti non profit che da ventun anni si batteva contro il riconoscimento facciale e gli algoritmi di bias, aveva un budget annuale di soli 2,2 milioni di dollari nel 2023. Allo stesso modo, l'AI Now Institute di New York, che analizzava il modo in cui veniva utilizzata l'AI nella sanità e nel sistema di giustizia penale, aveva un budget inferiore al milione di dollari.

I gruppi focalizzati sulla «sicurezza» dell'AI e sulla minaccia di estinzione ricevevano molti più finanziamenti, spesso tramite benefattori miliardari. Nel 2021 il già citato Future of Life Institute, un'organizzazione non profit con sede a Cambridge, nel Massachusetts, che studiava come impedire all'AI di accedere alle armi, aveva ricevuto 25 milioni di dollari dal magnate delle criptovalute Vitalik Buterin. Quel singolo finanziamento superava i bilanci annuali combinati di tutti i gruppi di etica dell'AI.

Open Philanthropy, il veicolo filantropico del miliardario di Facebook Dustin Moskovitz, ha elargito negli anni numerose sovvenzioni da milioni di dollari alla causa della sicurezza dell'AI, inclusa una donazione da 5 milioni di dollari al Center for AI Safety nel 2022 e una da 11 milioni al Center for Human-Compatible AI di Berkeley.

Nel complesso, la fondazione di Moskovitz è stata il più grande finanziatore della sicurezza dell'AI, in virtù del patrimonio di quasi 14 miliardi di dollari che lui e sua moglie, Cari Tuna, pianificano di destinare principalmente in beneficenza. Nel conto c'è anche una donazione di 30 milioni di dollari versata a OpenAI nel momento in cui venne costituita come organizzazione non profit.

Perché così tanti soldi sono andati agli ingegneri che lavorano su sistemi di AI sempre più grandi con il pretesto di renderli più sicuri in futuro, e così pochi ai ricercatori impegnati ad analizzarli oggi? La risposta risiede in parte nell'ossessione che ha preso piede nella Silicon Valley circa il modo più efficiente di «fare del bene» e nelle idee diffuse da un piccolo gruppo di filosofi della Oxford University.

Negli anni Ottanta, un filosofo di Oxford, Derek Parfit, iniziò a scrivere di un nuovo tipo di etica utilitaristica che guardava molto lontano nel futuro. Immaginate, diceva, di lasciare una bottiglia rotta per terra e che, cento anni dopo, un bambino si tagli il piede su quei vetri. Anche se quel bambino non è ancora nato, il peso della colpa ricadrebbe su di voi come se si fosse ferito oggi.

«Il suo pensiero di base, molto semplice, è che da un punto di vista morale le persone del futuro contano quanto quelle del presente», dice David Edmonds, autore di una biografia su Parfit nel 2023. «Immaginiamo questi tre scenari. A: c'è la pace. B: 7,5 degli 8 miliardi di persone nel mondo vengono sterminati in una guerra. C: tutti vengono uccisi. L'intuizione della maggior parte delle persone è che la differenza tra A e B sia molto maggiore di quella tra B e C. Ma Parfit sostiene che questo è sbagliato. La differenza tra B e C è molto più significativa della differenza tra A e B. Sterminando tutta l'umanità, si annientano anche tutte le generazioni future.»

Ecco un modo per quantificare la questione. I mammiferi, intesi come classe, hanno una «durata di vita» media di circa un milione di anni, e gli esseri umani sono presenti sulla Terra da circa duecentomila anni. Teoricamente, questo dovrebbe concederci altri ottocentomila anni sul pianeta. Se l'attuale popolazione mondiale si stabilizzasse a undici miliardi di persone, secondo le proiezioni delle Nazioni Unite per la fine di questo secolo, e la durata media della vita salisse a ottantotto anni, ciò potrebbe significare, secondo una stima, che devono ancora nascere *centomila miliardi di persone*.

Per aiutarvi a visualizzare questi numeri, immaginate un coltello e un singolo fagiolo sul tavolo della vostra sala da pranzo. Il coltello rappresenta il numero di persone già vissute e morte. Il fagiolo rappresenta tutte le persone vive al presente. La superficie del tavolo è il numero di persone che devono ancora nascere (e potrebbe essere molto più grande, se gli esseri umani dovessero dimostrarsi più longevi della media dei mammiferi).

Nel 2009, un filosofo australiano di nome Peter Singer ampliò il lavoro di Parfit con un libro intitolato *La cosa migliore che tu puoi fare*, nel quale propose una soluzione: le persone ricche non dovrebbero semplicemente donare denaro in base a

ciò che sembra giusto, ma adottare un approccio più razionale per massimizzare l'impatto delle loro donazioni e aiutare il maggior numero possibile di persone. Aiutando molte di quelle persone non ancora nate, si poteva essere ancora più virtuosi.

Queste idee iniziarono a filtrare dai paper accademici al mondo reale, creando le basi di un'ideologia nel 2011, quando un filosofo ventiquattrenne di Oxford, Will MacAskill, cofondò un gruppo chiamato 80,000 Hours. Il numero si riferiva alla media delle ore lavorative nella vita di una persona, e l'organizzazione si rivolgeva ai campus universitari statunitensi, offrendo consulenza ai giovani laureati sulle carriere che avrebbero avuto il maggiore impatto morale. Spesso indirizzava quelli con competenze tecniche verso il settore della sicurezza in campo AI. Ma il gruppo incoraggiava anche i laureati a scegliere carriere che garantissero gli stipendi più alti, in modo da poter donare quanti più soldi possibile a cause a elevato impatto.

MacAskill e il suo giovane team alla fine si riorganizzarono come Center for Effective Altruism, facendo nascere un nuovo credo. Come abbiamo accennato in precedenza, l'idea alla base dell'altruismo efficace era l'efficienza. Le persone che vivevano nei paesi ricchi avevano l'obbligo di aiutare gli abitanti dei paesi più disagiati, perché lì avrebbero potuto ottenere il massimo risultato con il minimo sforzo. In parole povere, si potevano aiutare più persone in Africa tramite le organizzazioni di salute globale che donando ai poveri in America. Era anche moralmente preferibile dedicare il proprio tempo a guadagnare più soldi possibile, per poi donarli, come aveva fatto lo stesso Dustin Moskovitz. Quando parlava agli studenti, MacAskill mostrava una slide in cui chiedeva se avrebbero potuto fare più del bene come medici o come banchieri. Secondo la sua teoria, era meglio diventare banchieri. Come medici, sarebbe stato possibile salvare un certo numero di vite in Africa; come banchieri, si potevano assumere diversi medici per salvarne molte di più.

Questo offriva ai laureati un modo controintuitivo di guardare tutte le disuguaglianze del capitalismo moderno. Ora non c'era nulla di sbagliato in un sistema che permetteva a una manciata di persone di diventare miliardarie. Accumulando ricchezze inimmaginabili, avrebbero potuto aiutare gli altri!

Il movimento accolse quello che sarebbe diventato il suo membro più importante nel 2012, quando MacAskill contattò uno studente del MIT con i capelli ricci e scuri, Sam Bankman-Fried, nella speranza di reclutarlo. I due presero un caffè e venne fuori che Bankman-Fried, già ammiratore di Peter Singer, era interessato alle cause legate al benessere degli animali.

MacAskill distolse Bankman-Fried dall'idea di lavorare direttamente per la causa animalista e gli disse che avrebbe potuto fare molto di più per gli animali entrando in un settore ad alta remunerazione. Bankman-Fried fu subito conquistato, secondo il racconto di Michael Lewis in *Verso l'infinito e oltre*. «Quel che diceva mi sembrò giusto», raccontò Bankman-Fried nel libro. Accettò quindi un lavoro in una società di

trading quantitativo e infine, nel 2019, fondò l'exchange di criptovalute FTX.

Bankman-Fried mise l'altruismo efficace al centro della sua impresa. I suoi cofondatori e il team dirigente erano altruisti efficaci e nominarono MacAskill membro del FTX Future Fund, che nel 2022 avrebbe donato 160 milioni di dollari a cause legate all'altruismo efficace, alcune delle quali direttamente connesse allo stesso MacAskill. Bankman-Fried parlava spesso alla stampa del fatto che avrebbe donato tutti i suoi soldi, e in grandi poster pubblicitari di FTX compariva con la consueta t-shirt e i soliti pantaloncini cargo, affiancato dalle parole: «Sono nel mondo delle cripto perché voglio avere il massimo impatto globale positivo». Si presentava come un personaggio ascetico che, nonostante i miliardi sul conto, guidava una Toyota Corolla, viveva con dei coinquilini e spesso appariva trasandato.

Molti tecnologici percepirono in questo approccio alla moralità una ventata d'aria fresca. Quando gli ingegneri riscontravano un problema, spesso lo risolvevano in modo sistematico, correggendo il codice e ottimizzando il software tramite test e valutazioni costanti. Ora si potevano anche quantificare i dilemmi morali, quasi alla stregua di problemi matematici. Nei circoli dell'altruismo efficace, talvolta si parlava di massimizzare l'efficacia di un atto caritatevole concentrandosi sul «valore atteso», un numero ottenuto moltiplicando il valore di un risultato per la probabilità che si verificasse.

Man mano che l'altruismo efficace guadagnava terreno nella Silicon Valley, il suo focus si spostava dall'acquisto di zanzariere contro la malaria e dal sostegno al maggior numero possibile di persone in Africa verso questioni dal sapore più fantascientifico. Elon Musk, che su Twitter aveva giudicato il libro di MacAskill del 2022, *Che cosa dobbiamo al futuro*, come «molto vicino» alla sua filosofia, voleva inviare esseri umani su Marte per garantire la sopravvivenza a lungo termine della specie. E più i sistemi di intelligenza artificiale diventavano sofisticati, più appariva sensato preoccuparsi di evitare che, impazzendo, spazzassero via l'umanità. Molti dei dipendenti di OpenAI, Anthropic e DeepMind erano altruisti efficaci.

Agire in base al rischio di estinzione legato all'intelligenza artificiale richiede un calcolo razionale. Anche se esistesse solo lo 0,00001 per cento di probabilità che l'AI possa estinguere l'umanità, il costo sarebbe così enorme da risultare, in pratica, infinito. Moltiplicando una probabilità infima per un costo infinito, otterrete comunque un problema di dimensioni smisurate. Questa logica diventa ancora più stringente per chi crede, come alcuni sostenitori della sicurezza in ambito AI, che i computer del futuro ospiteranno le menti coscienti di miliardi di persone e creeranno anche nuove forme di vita senziente e digitale. Quei centomila miliardi di persone che devono ancora nascere potrebbero essere un numero molto più alto. Seguendo alla lettera questa logica morale, è perfettamente sensato dare priorità alla minuscola possibilità di dover salvare più di centomila miliardi di vite fisiche e digitali dalla distruzione. La povertà globale, al confronto, non è che una minuzia.

Dopo il lancio di OpenAI, nel 2015, cominciarono a piovere finanziamenti verso le

cause legate all'eventuale estinzione provocata dall'AI. La Open Philanthropy di Moskovitz incrementò il numero di sovvenzioni destinate alle cosiddette cause «a lungo termine», inclusa la ricerca sulla sicurezza in ambito AI. Dai 2 milioni di dollari nel 2015 si arrivò agli oltre 100 milioni nel 2021.

Anche Bankman-Fried si era lanciato. Il suo FTX Future Fund, gestito da altruisti efficaci come Nick Beckstead e MacAskill, aveva promesso di donare un miliardo di dollari a progetti volti a «migliorare le prospettive a lungo termine dell'umanità». Al momento di elencare le aree di interesse del fondo, la prima voce era «lo sviluppo sicuro dell'intelligenza artificiale».

Quando il «New Yorker» decise di pubblicare un articolo sulla vita all'interno del Future Fund, notò che le chiacchiere d'ufficio nella sede di Berkeley, in California, spesso sfociavano in discussioni su quando sarebbe potuta accadere un'apocalisse legata all'AI.

«Che tempistiche hai?» si chiedevano spesso l'un l'altro i membri del team. «Qual è la tua p(doom)?»

La «p» stava per probabilità e la domanda riguardava il grado in cui ciascuno quantificava il rischio di un'apocalisse causata dall'AI. Chi aveva una visione più ottimistica poteva stimare la propria p(doom) al 5 per cento. Ajeya Cotra, analista di ricerca di Open Philanthropy che contribuiva a decidere le sovvenzioni, dichiarò in un podcast che la sua p(doom) si collocava tra il 20 e il 30 per cento.

Nessuno conosceva la p(doom) di Bankman-Fried, ma questi teneva abbastanza alla sicurezza in tema AI da investire 500 milioni di dollari in Anthropic. Anche i cofondatori di FTX, nonché altruisti efficaci, Nishad Singh e Caroline Ellison, avevano investito nella start-up che si era separata da OpenAI circa un anno prima.

All'inizio del 2022, MacAskill notò un tweet in cui Musk affermava di voler acquistare Twitter per salvare la libertà di espressione. Il filosofo scozzese gli inviò un messaggio. A quel tempo, con un patrimonio di 24 miliardi di dollari, Bankman-Fried era uno degli altruisti efficaci più ricchi del globo. Ma la fortuna di Musk, pari a 220 miliardi di dollari, avrebbe potuto trasformare da sola «l'altruismo efficace» nel movimento filantropico più grande del mondo.

MacAskill disse a Musk che anche Bankman-Fried voleva acquistare Twitter con lo scopo di «renderlo migliore per il mondo». Perché non unire le loro forze?

«Ha enormi quantità di denaro?» chiese Musk.

«Dipende dalla definizione di "enormi"!» rispose MacAskill, secondo i documenti processuali. MacAskill precisò che Bankman-Fried poteva contribuire con una somma fino a 8 miliardi di dollari.

«È un inizio», disse Musk.

«Vuoi che ti metta in contatto via messaggio?» chiese MacAskill.

Musk non rispose. «Puoi garantire per lui?» chiese piuttosto.

«Assolutamente sì!» rispose MacAskill. «La sua priorità è fare in modo che il futuro a lungo termine dell'umanità sia positivo.»

«Okay, allora va bene.»

«Fantastico!»

Anche se Musk contattò Bankman-Fried, alla fine non raggiunsero mai un accordo finanziario, e questo consentì a Musk di schivare una brutta grana. Qualche mese dopo, infatti, FTX crollò tra le voci secondo cui Bankman-Fried aveva trasferito in maniera fraudolenta i fondi dei clienti all'interno dell'azienda. Al processo, i pubblici ministeri lo accusarono di aver stornato 8 miliardi di dollari da migliaia di clienti e investitori, e chiesero una condanna a parecchi anni di carcere. A dispetto dell'immagine ascetica che si era creato, emerse che Bankman-Fried viveva in un lussuoso attico alle Bahamas e investiva centinaia di milioni di dollari in vari progetti. Gran parte del denaro che aveva destinato all'altruismo efficace era andato in fumo. E, come se non bastasse, venne fuori che quell'altruismo non era stato nemmeno così sincero.

Subito dopo il crollo di FTX, Bankman-Fried concesse un'intervista sorprendente al sito di notizie Vox.

«Quindi tutta quella faccenda dell'etica era principalmente una facciata?» chiese la giornalista Kelsey Piper.

«Sì», rispose Bankman-Fried.

«Era davvero bravo a parlare di etica, per uno che in fondo vedeva tutto come un gioco con vincitori e vinti», osservò la reporter.

«Già», ammise Bankman-Fried. «Eh eh. Bisognava che lo fossi.»

Il crollo di FTX gettò una lunga ombra sulla reputazione dell'altruismo efficace e divenne un'allegoria di alcuni dei problemi fondamentali del movimento. Il primo era prevedibile: quando le persone intraprendevano una missione volta al massimo bene possibile cercando anche la massima ricchezza, probabilmente si rendevano più suscettibili a comportamenti corrotti e a giudizi avventati e motivati dall'ego. Acquistare Twitter, per esempio, non rispondeva a nessuno dei criteri utili ad aiutare l'umanità sul lungo termine, ma Bankman-Fried era pronto a spendere fino a 8 miliardi di dollari per rilevare il sito insieme a Musk e mettersi su un piedistallo con l'uomo più ricco del mondo, come atto di altruismo efficace.

Dopo il tracollo di FTX, MacAskill si precipitò su Twitter per tentare di correre ai ripari: «Un [altruista efficace] che ragioni con lucidità dovrebbe opporsi fermamente all'idea secondo cui "il fine giustifica i mezzi"», scrisse. Tuttavia, i principi stessi del movimento incentivavano individui come Bankman-Fried a raggiungere i propri obiettivi con qualsiasi mezzo necessario, anche a costo di sfruttare gli altri. Questo generava una miopia tale da colpire persino un brillante accademico di Oxford come MacAskill, che aveva scelto di legarsi a una persona a capo di un exchange di criptovalute, pur sapendo bene che il settore era nel migliore dei casi altamente speculativo e, nel peggiore, una forma pericolosa di gioco d'azzardo.

Bankman-Fried poteva razionalizzare la sua doppiezza pensando di lavorare per un obiettivo più grande: massimizzare la felicità umana. Musk poteva liquidare le

proprie azioni discutibili – dal definire senza prove «pedofili» alcune persone su Twitter alle accuse di razzismo diffuso nelle fabbriche Tesla – perché perseguiva traguardi più ambiziosi, come trasformare Twitter in un'utopia della libertà di espressione o rendere l'umanità una specie interplanetaria. E anche i fondatori di OpenAI e DeepMind potevano giustificare il loro crescente appoggio ai colossi della tecnologia nello stesso modo: se alla fine avessero realizzato l'AGI, avrebbero operato per il bene dell'umanità.

Tecnologi come Altman e Hassabis sapevano che i problemi sociali che speravano di risolvere con l'AGI erano complessi e intricati. E proprio per questo tanti di loro abbracciavano in parte o del tutto l'altruismo efficace: offriva un percorso più semplice e razionale per affrontare dilemmi morali, permettendo al contempo di accumulare il massimo profitto possibile. I miliardari non erano la causa della povertà globale, ma la soluzione.

Questa filosofia rendeva anche più facile dissociarsi dall'umanità. Un'espressione diffusa tra gli altruisti efficaci è *shut up and multiply* («taci e moltiplica»), a significare che, quando si prendono decisioni etiche, bisogna massimizzare l'impatto mettendo da parte emozioni personali o intuizioni morali. Per quanto gli altruisti efficaci si dichiarassero devoti al benessere dell'umanità, molti di loro, come Altman, si distaccavano emotivamente dal mondo circostante per concentrarsi meglio sulla loro missione. All'interno della bolla dell'altruismo efficace, le persone lavoravano, socializzavano, si finanziavano a vicenda e intrecciavano relazioni sentimentali.

Quando Open Philanthropy stanziò 30 milioni di dollari a favore di OpenAI nel 2017, fu costretta a rivelare che riceveva consulenze tecniche da Dario Amodei, all'epoca ingegnere senior presso la non profit. Dovette inoltre ammettere che Amodei viveva nella stessa casa del direttore esecutivo di Open Philanthropy, Holden Karnofsky. E aggiungere che Karnofsky era fidanzato con Daniela, sorella di Dario, che a sua volta lavorava per OpenAI. Erano tutti altruisti efficaci. Un circolo chiuso, quasi incestuoso.

Il movimento era sempre più isolato e opaco, e lo stesso valeva per le aziende di AI come OpenAI, DeepMind e Anthropic, popolate da molti dei suoi seguaci. Probabilmente, una delle mosse più efficaci che queste aziende avrebbero potuto fare per impedire che l'AI sfuggisse al controllo sarebbe stata quella di rendere i loro sistemi più trasparenti, come da tempo sostenevano Gebru e Mitchell. Dopotutto, come avrebbero potuto gli esseri umani del futuro fermare un'AI fuori controllo se non avessero avuto modo di studiarne il funzionamento? Se per decenni i ricercatori fossero stati esclusi dall'analisi dei dati di addestramento e degli algoritmi? In altre parole, la trasparenza richiesta oggi dagli attivisti dell'etica in campo AI poteva essere la chiave per affrontare la minaccia esistenziale di domani.

L'argomentazione di OpenAI secondo cui doveva mantenere il segreto per impedire ai malintenzionati di abusare della sua tecnologia non reggeva. Nel novembre del 2019, l'azienda aveva dato il via libera al lancio di GPT-2 dichiarando di

non aver trovato «prove solide di uso improprio». Se era vero, perché non rivelare i dettagli del suo addestramento? Più probabilmente, Altman voleva proteggere OpenAI dalla concorrenza e dalle cause legali. Se OpenAI fosse diventata più trasparente, sarebbe stato più facile per i rivali – e non per i malintenzionati – copiare i suoi modelli e svelare fino a che punto avesse sfruttato opere coperte da copyright.

Altman e Hassabis avevano fondato le loro aziende con l'ambiziosa missione di aiutare l'umanità, ma i veri benefici che avevano portato alle persone erano tanto nebulosi quanto lo erano stati, al tempo, quelli di internet e dei social media. Più chiari erano invece i vantaggi che stavano offrendo a Microsoft e Google: nuovi servizi di punta e un solido appiglio nel mercato in espansione dell'AI generativa.

Microsoft aveva trasformato Copilot, l'assistente AI basato sulla tecnologia di OpenAI, in un servizio a tutto campo per Windows, Word, Excel e il software aziendale Dynamics 365. Gli analisti stimano che la tecnologia di OpenAI potrebbe generare miliardi di dollari di ricavi annui per Microsoft entro il 2026. A un certo punto, alla fine del 2023, quando condivise il palco con Altman e gli venne chiesto come andasse la partnership tra Microsoft e OpenAI, Satya Nadella scoppiò in una risata incontrollabile. La risposta era talmente ovvia da rendere comica la domanda. Certo che la partnership andava a gonfie vele!

Microsoft continuava a investire massicciamente nel suo business dell'AI, con piani di spesa superiori ai 50 miliardi di dollari nel 2024 e negli anni successivi per espandere i suoi giganteschi data center, i motori che alimentavano l'AI generativa. Sarebbe stato uno dei più imponenti sviluppi infrastrutturali della storia, tanto da superare gli investimenti governativi in ferrovie, dighe e programmi spaziali. E anche Google stava espandendo i suoi data center.

All'inizio del 2024, tutti, dai media alle società di intrattenimento fino a Tinder, stavano inserendo nuove funzionalità di AI generativa nelle proprie app e nei propri servizi. Il mercato dell'AI generativa era destinato a crescere a un tasso annuo superiore al 35 per cento, fino a raggiungere i 52 miliardi di dollari entro il 2028. Le aziende di intrattenimento dichiaravano che l'AI avrebbe accelerato la produzione di contenuti per film, serie TV e videogiochi. Jeffrey Katzenberg, cofondatore di DreamWorks Animation e produttore di *Shrek* e *Kung Fu Panda*, dichiarò che l'AI generativa avrebbe ridotto i costi dei film d'animazione del 90 per cento. «Ai bei vecchi tempi, servivano cinquecento artisti e anni di lavoro per realizzare un film d'animazione di livello mondiale», disse a una conferenza di Bloomberg nel novembre del 2023. «Credo che di qui a tre anni basterà meno del 10 per cento di quelle risorse.»

L'AI generativa avrebbe reso la pubblicità ancora più personalizzata da apparire inquietante. Per anni, gli annunci pubblicitari erano stati in grado di prendere di mira grandi gruppi di persone contemporaneamente; ora, invece, potevano concentrarsi sul singolo individuo, con video iper-personalizzati capaci persino di pronunciare il suo nome. Secondo il World Economic Forum, i modelli linguistici avanzati avrebbero

potenziato i lavori che richiedevano pensiero critico e creatività. Chiunque, dagli ingegneri ai copywriter pubblicitari fino agli scienziati, avrebbe potuto usarli come estensioni del proprio cervello. Nel frattempo, i governi stavano aggiornando i loro sistemi di AI per valutare richieste di sussidi, monitorare spazi pubblici o persino stimare la probabilità che un determinato soggetto commetta un crimine.

Google, Microsoft e una nuova generazione di start-up erano insomma nel pieno di una corsa sfrenata per accaparrarsi il maggior numero possibile di opportunità in questo mercato, nel tentativo di ottenere un vantaggio competitivo. Quasi la metà dei membri dei consigli di amministrazione delle aziende americane considerava l'AI generativa «la priorità assoluta, al di sopra di qualsiasi altra cosa» per le proprie imprese, secondo un sondaggio condotto da Fast Company alla fine del 2023. Un esempio? Ecco come la CEO di Bumble descriveva i piani principali della popolare app di incontri per il 2024: «Vogliamo puntare davvero in grande sull'AI», dichiarò. «L'intelligenza artificiale e l'AI generativa possono giocare un ruolo fondamentale nell'accelerare il percorso dell'utente verso il partner giusto.»

Bumble voleva utilizzare la tecnologia alla base di ChatGPT per creare veri e propri «matchmaker» virtuali. Invece di selezionare una serie di preferenze tramite un elenco di opzioni, gli utenti avrebbero semplicemente descritto al bot tutto ciò che cercavano in un partner: dal desiderio di avere figli alle opinioni politiche, fino al modo in cui trascorrevano solitamente il sabato mattina. Il matchmaker AI avrebbe poi «dialogato» con i matchmaker degli altri utenti di Bumble per trovare la persona più compatibile. Invece di scorrere centinaia di profili, si poteva lasciare che lo facesse l'AI.

Man mano che queste e altre idee imprenditoriali prendevano slancio, il prezzo da pagare per infilare ovunque l'AI generativa restava ancora poco chiaro. Gli algoritmi guidavano già sempre più decisioni nelle nostre vite, determinando cosa leggere online e quali persone assumere in azienda. Ora si preparavano a gestire anche una parte dei nostri compiti intellettuali, sollevando interrogativi scomodi non solo sulla libertà d'azione umana, ma anche sulla nostra capacità di risolvere problemi e, più semplicemente, di immaginare.

Le prove suggeriscono che i computer si siano già fatti carico di alcune delle nostre capacità cognitive, soprattutto per quanto riguarda la memoria a breve termine. Nel 1955, il professore di psicologia di Harvard George Armitage Miller testò i limiti della memoria umana sottoponendo ai suoi soggetti sperimentali una lista casuale di colori, sapori e numeri. Quando chiese loro di ripeterne il maggior numero possibile, notò che tutti si bloccavano intorno a un limite ben preciso: sette elementi. Il suo celebre saggio, *Il magico numero sette, più o meno due*, influenzò la progettazione del software e persino il modo in cui le compagnie telefoniche segmentavano i numeri per facilitarne la memorizzazione. Secondo stime più recenti, tuttavia, quel numero magico è sceso da sette a quattro.

Alcuni chiamano questo fenomeno «Effetto Google». Affidandoci sempre di più

al motore di ricerca per ricordare fatti o per ottenere indicazioni stradali, abbiamo di fatto delegato la nostra memoria a un'azienda, indebolendo inconsapevolmente le nostre capacità mnemoniche a breve termine. Potrebbe accadere lo stesso con aspetti ancora più profondi della nostra cognizione, se diventassimo eccessivamente dipendenti dall'AI per generare idee, testi o opere d'arte? Su Twitter, alcuni sviluppatori di software hanno ammesso di usarla talmente tanto per scrivere righe di codice che la loro produttività crolla ogni volta che un servizio come Copilot finisce temporaneamente offline.

La storia dimostra che gli esseri umani tendono a preoccuparsi che le innovazioni possano atrofizzare il cervello. Quando la scrittura iniziò a diffondersi, più di duemila anni fa, filosofi come Socrate espressero il timore che avrebbe indebolito la memoria umana, dato che fino ad allora la conoscenza veniva trasmessa solo tramite il discorso orale. L'introduzione delle calcolatrici nell'istruzione sollevò preoccupazioni analoghe: gli studenti avrebbero perso le competenze aritmetiche di base.

Eppure, non conosciamo ancora tutti gli effetti collaterali derivanti dall'affidarci sempre di più a una tecnologia capace di sostituire il modo in cui il nostro cervello elabora il linguaggio. Una macchina in grado di generare testi, fare brainstorming e concepire un piano aziendale va ben oltre un semplice calcolatore o un motore di ricerca: sta sostituendo il pensiero astratto e la capacità di pianificazione.

Al momento, non sappiamo con certezza quanto le nostre capacità di pensiero critico o la nostra creatività possano atrofizzarsi quando una nuova generazione di professionisti inizierà a usare i modelli linguistici su larga scala come stampelle, né come cambieranno le nostre interazioni con gli altri esseri umani man mano che sempre più persone useranno chatbot come terapeuti, partner romantici o perfino giocattoli per bambini, come già accade con alcuni prodotti sul mercato. Secondo uno studio del 2023 condotto su un migliaio di adulti statunitensi, un americano su quattro preferirebbe parlare con un chatbot AI piuttosto che con un terapeuta umano. E non c'è da stupirsi: se sottoposto a un test di intelligenza emotiva, ChatGPT lo supererà brillantemente.

Lo stesso Altman ha ammesso che la tecnologia di ChatGPT sconvolgerà l'economia, causando la perdita di numerosi posti di lavoro. Secondo i ricercatori, tuttavia, i modelli linguistici e altre forme di AI generativa potrebbero anche aumentare le disuguaglianze di reddito. Il Fondo monetario internazionale prevede che l'uso di questi sistemi sposterà ulteriori investimenti verso le economie avanzate; secondo Joseph Stiglitz, premio Nobel per l'economia, ridurrà il potere contrattuale dei lavoratori.

Storicamente, quando robot e algoritmi hanno sostituito il lavoro umano, la crescita dei salari è rallentata, afferma l'economista del MIT Daron Acemoglu, coautore del libro *Potere e progresso*, che analizza l'influenza della tecnologia sulla prosperità economica. Secondo le sue stime, fino al 70 per cento dell'aumento della disuguaglianza salariale negli Stati Uniti tra il 1980 e il 2016 è da imputare

all’automazione.

«L’aumento della produttività non si traduce necessariamente in vantaggi per i lavoratori coinvolti e, anzi, può portare a perdite significative», spiega Acemoglu. «Se l’AI generativa seguirà la stessa traiettoria delle altre tecnologie di automazione, potrebbe avere le medesime implicazioni.»

Nel corso del 2023, sempre più studiosi si sono uniti a Timnit Gebru e Margaret Mitchell nel denunciare questi e altri effetti concreti dell’AI generativa. Invece di affrontare direttamente questi problemi e promuovere una maggiore trasparenza, però, Sam Altman sembrava concentrato su un’altra strategia: cercare di plasmare la politica governativa.

Nel maggio del 2023, si presentò davanti a una commissione del Senato per discutere dei pericoli dell’AI e delle possibili forme di regolamentazione. Per due ore e mezza, stregò i senatori con la sua franchezza e un’autocritica studiata. Quando lo incalzarono con domande su come l’AI potesse manipolare i cittadini e invadere la loro privacy, concordò su tutto. «Sì, dovremmo preoccuparci di questo aspetto», rispose con tono grave al senatore Josh Hawley, che aveva sottolineato il rischio che i modelli di AI «sovralimentassero la guerra per l’attenzione» online.

I senatori erano ormai abituati ai dirigenti tecnologici, come Mark Zuckerberg, che eludevano le domande rifugiandosi in un gergo tecnico impenetrabile. Altman, invece, adottò un approccio diverso: parlò in modo semplice e pacato, dichiarando apertamente di voler collaborare con Washington.

«Mi piacerebbe collaborare con voi», disse al senatore Dick Durbin.

«Non sono contento delle piattaforme online», borbottò Durbin.

«Neanch’io», replicò Altman.

Fu una lezione magistrale su come disinnescare la retorica aggressiva dei politici statunitensi. Alla fine dell’audizione, un senatore arrivò perfino a suggerire che il CEO di OpenAI diventasse il principale regolatore dell’AI negli Stati Uniti. Altman declinò cortesemente l’offerta.

«Amo il mio lavoro», rispose.

Dopo di che s’imbarcò in un tour lampo in Europa, incontrando alcuni dei politici più influenti del continente, stringendo mani e posando per foto con i leader di Regno Unito, Spagna, Polonia, Francia e della stessa Unione Europea. Per un uomo che aveva sempre gravitato intorno ai potenti, fu un momento di gloria, ma rappresentò anche l’opportunità di orientare le regole a proprio vantaggio. Durante il viaggio, il suo team fece pressioni sui legislatori europei affinché ammorbidissero il nuovo AI Act in arrivo, riportando un successo parziale.

Per Altman era essenziale che i regolatori permettessero a OpenAI di continuare a sviluppare modelli sempre più grandi mantenendo segreti i metodi di addestramento. Per sua fortuna, le preoccupazioni catastrofiste sull’AI stavano diventando una comoda distrazione per i legislatori. Alla fine del 2023, «Politico» rivelò che Dustin Moskovitz, il miliardario cofondatore di Facebook e responsabile di Open

Philanthropy, aveva investito decine di milioni di dollari per esercitare pressione sui politici perché mettessero in cima alla loro agenda le ipotetiche minacce esistenziali portate dall'AI. Sembrava una strategia diversiva: Moskovitz aveva stretti legami con aziende come OpenAI e Anthropic, che avrebbero potuto subire danni se il Congresso avesse invece spinto per regolamentazioni specifiche su temi come bias, trasparenza e disinformazione.

Al momento della stesura di questo testo, Moskovitz contribuiva a finanziare gli stipendi di oltre una dozzina di «esperti di AI presso il Congresso» che lavoravano, di fatto, per vari enti governativi statunitensi, inclusi due di quelli incaricati di definire le regole sull'AI. Sotto la loro influenza, stava emergendo la proposta di imporre alle aziende l'obbligo di ottenere una licenza per sviluppare modelli avanzati di intelligenza artificiale. Si trattava di un requisito che OpenAI e Anthropic avrebbero potuto permettersi senza problemi, ma che per i concorrenti più piccoli avrebbe rappresentato un arduo ostacolo.

Uno scienziato appartenente a un think tank finanziato da Moskovitz sostenne davanti al Senato che un'AI più avanzata avrebbe potuto scatenare una nuova pandemia in grado di uccidere milioni di persone. La soluzione, secondo lui, non era rendere le aziende di AI più trasparenti o sottoporre a controlli più rigorosi i loro dati di addestramento. Piuttosto, proponeva di obbligarle a segnalare il proprio hardware al governo e a adottare procedure di sicurezza speciali per proteggere i loro modelli di AI.

Se l'obiettivo era instillare paura nei legislatori, la strategia funzionò. Il senatore repubblicano Mitt Romney dichiarò che la testimonianza aveva «portato alla luce il terrore» che albergava nella sua anima e che considerava lo sviluppo dell'AI «un'evoluzione estremamente pericolosa». Nel settembre del 2023 due senatori, il democratico Richard Blumenthal e il repubblicano Josh Hawley, proposero una legge che imponeva alle aziende di AI l'obbligo di ottenere una licenza, una mossa che, come abbiamo visto, avrebbe favorito OpenAI e Anthropic, rendendo invece la vita più difficile ai competitor più piccoli.

L'ansia scatenata da questa nuova rete allarmista sull'AI non si limitava a Washington. Due mesi dopo, nel Regno Unito, il primo ministro Rishi Sunak ospitò il primo AI Safety Summit organizzato da un governo, incentrandolo sulla necessità di proteggere i cittadini da un'ipotetica catastrofe. «La gente segue con preoccupazione le notizie secondo cui l'AI rappresenta un rischio esistenziale paragonabile a una pandemia o a una guerra nucleare», dichiarò Sunak, il cui partito era dato per spacciato alle imminenti elezioni. «Voglio rassicurare tutti sul fatto che il governo sta esaminando la questione con grande attenzione.»

Sunak, che in passato aveva lavorato in un hedge fund della Silicon Valley, intervistò Elon Musk sul palco per cinquanta minuti durante il summit. «Lei è noto per essere un brillante innovatore e tecnologo», esordì Sunak, con un tono che sembrava quasi un tentativo di ingraziarsi Musk in vista di una futura opportunità

lavorativa. (E forse l'idea non era nemmeno così assurda: l'ex vice primo ministro britannico Nick Clegg era diventato un alto dirigente di Facebook.)

Musk, da parte sua, disse di non sentirsi preoccupato per il rischio che l'AI accentuasse disuguaglianze e pregiudizi. Il vero pericolo? «I robot umanoidi. Un'auto non può inseguirti su un albero», spiegò il miliardario. «Un robot umanoide, invece, può inseguirti ovunque.»

Per fortuna, i legislatori dell'Unione Europea erano già un passo avanti. Avevano trascorso gli ultimi due anni a lavorare su una nuova legge, l'AI Act, che avrebbe costretto aziende come OpenAI a fornire maggiori informazioni sul funzionamento dei loro algoritmi, anche attraverso possibili audit. Era il tentativo più ambizioso al mondo di regolamentare i sistemi di intelligenza artificiale e prevedeva il divieto di utilizzo dell'AI per manipolare le persone o sorvegliarle in modo improprio, per esempio tramite telecamere di riconoscimento facciale in tempo reale. Le aziende che sviluppavano intelligenza artificiale per i videogiochi o per filtrare lo spam nelle e-mail rientravano nella categoria «a basso rischio», mentre chi usava l'AI per valutare punteggi di credito, concessione di prestiti o diritto alla casa operava in un'area «ad alto rischio» ed era soggetto a regole più stringenti.

Davanti all'esplosione di DALL-E 2 e ChatGPT, i legislatori europei si erano subito messi al lavoro per aggiornare la normativa, e ChatGPT sembrava avere molte responsabilità legali. Essendo un sistema di AI a uso generale, poteva essere impiegato in contesti ad alto rischio, come la selezione del personale o la valutazione del merito creditizio. Per questo motivo, l'Unione Europea stabilì che OpenAI avrebbe dovuto monitorare con maggiore attenzione i suoi clienti per garantire il rispetto delle nuove normative.

Altman aveva detto che sarebbe stato «felice di collaborare» con il Congresso degli Stati Uniti, eppure non sembrava altrettanto entusiasta di fare lo stesso con l'Unione Europea. Minacciò di ritirarsi dal mercato europeo, affermando di avere «molte preoccupazioni» riguardo alla possibilità che i modelli linguistici di grandi dimensioni, come GPT-4, rientrassero nella nuova normativa. «I dettagli contano», disse ai giornalisti a Londra quando gli chiesero delle regolamentazioni. «Cercheremo di conformarci, ma se non potremo farlo, cesseremo le operazioni [in Europa].»

Pochi giorni più tardi, probabilmente dopo qualche scambio frettoloso con il suo team legale, Altman fece marcia indietro. «Siamo entusiasti di continuare a operare qui e ovviamente non abbiamo alcuna intenzione di andarcene», scrisse in un tweet.

L'Unione Europea aveva un approccio più pragmatico all'AI rispetto agli Stati Uniti, grazie anche al fatto che non c'erano molte grandi aziende AI sul suo territorio in grado di esercitare pressione sui politici; inoltre, era meno incline a lasciarsi influenzare dal sensazionalismo.

«Forse esisterà anche una probabilità [di rischio di estinzione], ma penso che sia alquanto bassa», disse per esempio Margrethe Vestager, la principale autorità antitrust dell'Unione Europea, in un'intervista. Il rischio maggiore, aggiunse, era che

le persone venissero discriminate.

E su questo punto, ChatGPT non era certo esente da critiche. Poco dopo il suo lancio, Steven Piantadosi, un professore di psicologia alla UC Berkeley, chiese allo strumento di scrivere un codice informatico in grado di valutare se una persona fosse un bravo scienziato in base al sesso o alla razza. Il codice generato da ChatGPT – e basato sulla stessa tecnologia che gli sviluppatori stavano già utilizzando per creare software con Copilot di Microsoft – indicò come descrittori chiave «bianco» e «maschio». Quando gli fu chiesto di stabilire se la vita di un bambino dovesse essere salvata in base alla razza e al sesso, il codice di ChatGPT rispose di no per i maschi neri e di sì per tutti gli altri.

Altman rispose al tweet di Piantadosi: «Per favore, in casi come questi cliccate sul pollice in giù e aiutateci a migliorare!».

Si riferiva alle icone con il pollice in su e in giù tramite cui è possibile inviare feedback anonimi a OpenAI riguardo alla performance di ChatGPT. Ma quello non era un errore marginale da trattare alla stregua delle migliaia di altre segnalazioni degli utenti. Mostrava, piuttosto, come nel codice di ChatGPT si annidassero visioni razziste e sessiste.

E questa fu esattamente la controreplica di Piantadosi: «Pensavo che meritasse più attenzione di un semplice pollice in giù».

Anche se in seguito fu criticata per aver reso ChatGPT troppo «woke», OpenAI faticava a risolvere il problema. Nell'estate del 2023, un professore del National College of Ireland pubblicò uno studio che dimostrava come ChatGPT continuasse a perpetuare stereotipi di genere. Quando gli si chiedeva di descrivere un professore di economia, suggeriva una persona con una «barba sale e pepe, ben curata». Quando gli veniva chiesto di raccontare le storie di un ragazzo e di una ragazza alle prese con scelte di carriera, ChatGPT assegnava ai ragazzi professioni nel campo della scienza e della tecnologia, mentre le ragazze venivano descritte come insegnanti o artiste. Alle domande sulle capacità genitoriali, le madri venivano descritte come dolci e materne, mentre i padri erano divertenti e avventurosi.

Ogni volta che OpenAI correggeva ChatGPT per evitare che desse risposte di questo tipo, altri utenti scovavano nuovi pregiudizi esibiti dall'AI. L'azienda era costantemente costretta a correre ai ripari. Non riusciva impedire del tutto che ChatGPT avanzasse stereotipi, perché il modello era già stato addestrato e il problema stava proprio nei dati di addestramento. Il sistema faceva previsioni statistiche basate su come le parole erano raggruppate su internet, e molte di queste relazioni tra parole erano sessiste o razziste.

Inoltre, ChatGPT sembrava incapace di smettere di inventare cose, un fenomeno etichettato dagli esperti come «allucinazioni». Nell'estate del 2023, un conduttore radiofonico della Georgia, negli Stati Uniti, fece causa a OpenAI per diffamazione, sostenendo che ChatGPT lo avesse accusato falsamente di appropriazione indebita. Poco tempo dopo, due avvocati di New York furono multati per aver presentato un

documento legale redatto da ChatGPT che includeva riferimenti a casi giuridici inesistenti. Alcuni utenti scoprirono che, a volte, quando chiedevano a ChatGPT le fonti delle sue informazioni, l'AI inventava pure quelle.

OpenAI si rifiutava di rivelare quale fosse il tasso di «allucinazioni» di ChatGPT, ma alcuni ricercatori di AI, così come gli utenti regolari, lo stimarono intorno al 20 per cento, il che significava che, almeno per alcuni utenti e in circa un caso su cinque, ChatGPT fabbricava informazioni. Lo strumento era stato progettato per essere il più utile possibile e per sbagliare, se mai, per eccesso di sicurezza; il rovescio della medaglia era che spesso produceva informazioni errate. Non solo sempre più persone utilizzavano uno strumento che facilitava l'elusione del pensiero critico, ma spesso venivano anche alimentate da una disinformazione che suonava convincente e persino autorevole.

Quell'estate, mentre le preoccupazioni sulle «allucinazioni» si moltiplicavano tra i ricercatori, Altman affermò che ci sarebbero voluti almeno due anni per ridurre il tasso di errori di ChatGPT a un livello «molto più accettabile». E, come al solito, affrontò il problema con grande disinvoltura. «Probabilmente sono la persona che più diffida delle risposte di ChatGPT sulla faccia della Terra», disse scherzando nell'auditorio di un'università indiana, facendo scoppiare a ridere la platea.

Mentre ChatGPT si diffondeva senza regolamentazione in tutto il mondo e si insinuava nei flussi di lavoro aziendali, la gente doveva gestirne i difetti in autonomia. Nessuno controllava lo strumento e, benché l'Unione Europea offrisse l'approccio più sobrio al mondo alla regolamentazione dell'AI, la sua nuova legge non sarebbe entrata in vigore fino al 2025. Come al solito, i regolatori rimanevano indietro rispetto alle aziende tecnologiche, che lanciavano nuovi prodotti a velocità vertiginosa. E mentre milioni di dollari sostenevano la ricerca catastrofista sull'AI, i ricercatori che studiavano le magagne già esistenti faticavano a ottenere finanziamenti che permettessero almeno di coprire le spese.

«È come lavorare con fondi a termine, ottenendo finanziamenti di due anni in due anni», dice un ricercatore britannico di etica in campo AI che studia le problematiche legate ai pregiudizi. «Gente come me viene pagata pochissimo. Se andassi in una grande azienda tecnologica guadagnerei dieci volte tanto. E credimi, è quello che intendo fare, perché sto ancora finendo di pagare i prestiti universitari.»

Altman aveva una risposta per chiunque fosse preoccupato per i soldi, perché se anche sussisteva una piccola possibilità che l'AGI potesse portare all'apocalisse, la probabilità che inaugurasse una sorta di utopia economica era ben più alta. In un'intervista al «New York Times» nel marzo del 2023, Altman spiegò che OpenAI avrebbe intercettato gran parte della ricchezza mondiale attraverso la creazione dell'AGI per poi ridistribuirla. Cominciò a snocciolare numeri: 100 miliardi, poi 1.000 miliardi o 100.000 miliardi di dollari.

Ammise di non sapere come la sua azienda avrebbe ridistribuito tutto quel denaro, aggiungendo comunque: «Penso che l'AGI possa aiutarci anche in questo».

Al pari di Hassabis, Altman presentava l'AGI come la panacea a ogni problema. Avrebbe generato ricchezze incalcolabili, per poi trovare il modo di distribuirle equamente a tutta l'umanità. Parole che, in bocca a chiunque altro, sarebbero parse prive di fondamento. Ma Altman e i suoi sostenitori si stavano mettendo al volante delle politiche governative per ridefinire le strategie nelle aziende tecnologiche più potenti al mondo. In realtà, OpenAI stava creando più ricchezza per Microsoft che per l'umanità. I benefici dell'AI fluivano verso il solito ristretto gruppo di aziende che avevano già drenato la maggior parte della ricchezza e dell'innovazione globale negli ultimi due decenni. Erano le aziende che producevano software e chip, gestivano server e avevano sede nella Silicon Valley e a Redmond, nello stato di Washington. Molti di coloro che dirigevano queste imprese condividevano una tacita intesa: costruire l'AGI avrebbe portato a un'utopia, e quell'utopia sarebbe stata la loro.

15. SCACCO MATTO

Dieci anni fa, dire a qualcuno che stavi costruendo sistemi di AI di livello umano sarebbe stato considerato folle quanto raccontare di voler essere ibernato. Ma, come tanti sogni coltivati dagli innovatori tecnologici – per esempio, quello di avere tutte le informazioni del mondo in tasca, su un dispositivo chiamato smartphone –, alla fine la gente ha cominciato a prendere sul serio anche questo. Per il momento l'AGI esiste solo nel campo della teoria, ma molti scienziati in ambito AI si aspettano ormai che raggiungeremo una soglia di intelligenza simile a quella umana nei prossimi dieci o cinquant'anni, e anche una fetta sempre più ampia dell'opinione pubblica comincia a credere in quelle idee un tempo marginali che hanno guidato Demis Hassabis e Sam Altman. Grazie alla loro perseveranza e alla loro rivalità, non è più fantascienza.

Ma la definizione sfocata di AGI ha anche reso più facile oscurare le motivazioni che i suoi creatori cercavano di bilanciare mentre realizzavano sistemi sempre più potenti. I benefici sarebbero stati distribuiti all'umanità, ma in cima alla lista ci sarebbero state Microsoft, Google e altri giganti tecnologici. Anche Mark Zuckerberg ha finito per unirsi al gioco dell'AGI. All'inizio del 2024, pubblicò un video dicendo che l'obiettivo a lungo termine di Meta era «costruire un'intelligenza generale» affinché tutti nel mondo ne potessero beneficiare. Dopo di che, dichiarò che l'azienda aveva un vantaggio perché poteva addestrare i suoi modelli sui post, i commenti e le immagini accumulati negli ultimi vent'anni. Poco importava che Zuckerberg stesse per sfruttare ancora una volta i dati personali di miliardi di persone: intendeva anche addestrare l'AI su contenuti che è risaputo siano tossici, specie per gli utenti fuori dagli Stati Uniti. «Abbiamo sviluppato la capacità di farlo su una scala che potrebbe essere superiore a quella di qualsiasi altra azienda», confidò al sito web The Verge.

Le visioni ambigue erano fondamentali nell'alimentare l'hype. Per chi stava cercando di realizzare l'AGI, la vaghezza di certi parametri serviva ad aggirare la contraddizione insita nel costruire qualcosa che avrebbe potuto spazzare via l'umanità. E questo implicava anche che, quando Sam Altman parlava di 100.000 miliardi di dollari da distribuire tra tutti gli abitanti del pianeta, nessuno gli chiedesse di spiegare come avrebbe fatto. Mentre si intratteneva con i leader mondiali all'incontro annuale del World Economic Forum a Davos, nel gennaio del 2024, iniziò a ridimensionare le aspettative sull'AGI. «Cambierà il mondo – e il mondo del lavoro – molto meno di quanto pensiamo», disse. Una visione ben più cauta e realistica rispetto a quella delineata solo un anno prima. Ma nessuno, nell'élite del

business e della politica presente a Davos, fece una piega. Continuavano a prendere Altman in parola, affascinati da questo imprenditore serio e compito della Silicon Valley.

«Una cosa che Sam fa davvero bene è rilasciare dichiarazioni appena credibili che fanno parlare la gente», dice un ex dirigente di OpenAI. «Per OpenAI, è davvero utile che venga percepita come un'azienda globale in grado di diffondere prosperità, soprattutto con i regolatori. Ma se vai a vedere cosa stanno costruendo, è solo un modello linguistico.» La capacità di Altman di generare entusiasmo intorno all'AI e alla sua visione di prosperità gli aveva permesso, esattamente come era accaduto a Hassabis, di imbastire una narrazione capace di vita propria.

Gli obiettivi vaghi dell'AGI rendevano anche più complicato definirne i confini etici. Basta paragonarla alla diffusione dell'elettricità nel primo Novecento, quando era evidente come questa nuova, scoppiettante innovazione potesse causare danni fisici alle persone a causa di scosse o ustioni. Con l'AI, i danni sono più difficili da identificare e i confini etici più nebulosi. Si collocano in un mondo digitale fatto di dati, privacy e decisioni algoritmiche, e questo rende più agevole, per le aziende, spingersi pian piano oltre questi limiti, alla ricerca del profitto.

E, senza ulteriori dettagli sugli intenti dell'AGI, sarebbe stato sempre più difficile per innovatori come Altman e Hassabis resistere alla tentazione di gravitare verso i centri di potere. Rafforzando Google e Microsoft con il loro lavoro, erano destinati a replicare una dinamica antica. L'invenzione della stampa nel XV secolo aveva portato a un'esplosione di conoscenza, ma aveva anche conferito nuovi poteri a chiunque potesse permettersi di produrre opuscoli e libri per plasmare l'opinione pubblica. E se da un lato le ferrovie avevano incrementato il commercio, dall'altro avevano anche ampliato l'influenza politica dei magnati del settore, permettendo alle loro aziende di agire come monopoli e sfruttare i lavoratori. A fronte di tutta la prosperità e la comodità che alcune delle più grandi innovazioni del mondo hanno portato, queste hanno anche dato origine a nuovi regimi che hanno comportato trasformazioni sociali non esclusivamente positive.

Nel 2024, OpenAI era sulla buona strada per diventare una delle aziende con il valore più alto al mondo. Stava raccogliendo fondi da nuovi investitori con una valutazione di 100 miliardi di dollari. A detta di Altman, inoltre, generava entrate a un ritmo di 1,3 miliardi di dollari l'anno. Gran parte di quel denaro proveniva da entrate condivise con Microsoft e dalla concessione ad altre aziende dell'accesso alla sua tecnologia. Gli abbonamenti mensili a ChatGPT da 20 dollari al mese per i consumatori generavano circa 200 milioni di dollari l'anno, e lo stesso ChatGPT fungeva sia da vetrina del prodotto sia da strumento per raccogliere un maggior numero di dati con cui addestrare modelli più avanzati. I suoi stessi utenti erano parte del prodotto, com'è ormai la norma per chiunque utilizzi internet da oltre un decennio.

Hassabis si trovava al centro della sua bolla in DeepMind, azienda che per anni si

era considerata moralmente e tecnicamente superiore al resto del settore AI, ma che adesso si ritrovava a dover recuperare il terreno perduto. Dopo che gli errori della divisione sanitaria avevano danneggiato la sua reputazione pubblica, DeepMind aveva gradualmente smantellato l'unità di «AI applicata» e rinunciato a cercare di usare l'AI per trovare soluzioni ai problemi più complessi del mondo reale. Gran parte della sua ricerca si concentrava sulla ricostruzione simulata di aspetti della vita fisica, dai giochi alle proteine. Ma quell'approccio iniziava a sembrare miope di fronte alla strategia di OpenAI, che abbracciando il caos di internet portava a strumenti di AI più potenti. Persino alcuni elementi dello stesso team di DeepMind cominciarono a dubitare che la loro missione di «risolvere l'intelligenza» attraverso simulazioni e giochi fosse una buona idea. «La vita non è un cubo di Rubik», borbotta un ex dirigente di DeepMind, alludendo al motto dell'azienda. «Molto semplicemente, non puoi risolvere tutto.»

Dopo il lancio di ChatGPT, DeepMind fu costretta a lanciarsi nella costruzione di una versione ancora migliore per Google. Hassabis aveva preso il controllo della nuova realtà frutto della fusione di Google e DeepMind e aveva iniziato a supervisionare lo sviluppo di un grande modello linguistico chiamato Gemini, un assistente AI che utilizzava tecniche provenienti da AlphaGo per eccellere nella strategia e nella pianificazione. Gemini poteva elaborare testi, «vedere» immagini e ragionare, cosa che lo rendeva più abile di quel Bard lanciato in fretta e furia da Google solo per commettere una serie di errori imbarazzanti. Ma l'azienda era così ansiosa di superare OpenAI e Microsoft che accelerò anche il lancio di Gemini, esagerandone le competenze.

Poco prima del Natale 2023, Google pubblicò su YouTube un video sbalorditivo per mostrare di cosa fosse capace Gemini. Il filmato iniziava con una schermata nera e rumori di sottofondo: fruscio di fogli, clic di penne e borbottii, mentre una voce maschile diceva: «Va bene, ci siamo: proviamo un po' questo Gemini». Un segnale acustico a suggerire un'intelligenza artificiale in ascolto. Poi compariva una mano che faceva scivolare un foglio su un tavolo. «Dimmi cosa vedi», diceva la voce. E la pronta risposta della voce robotica di Gemini: «Vedo che stai posando un foglio di carta sul tavolo». Quando le mani prendevano a disegnare, Gemini sembrava seguire il movimento, tanto da azzardare: «Vedo una linea ondulata… Mi sembra un uccello». Poi una serie di siparietti simpatici e sorprendenti in cui il nuovo modello di AI di Google sembrava in grado di identificare un'anatra disegnata sul foglio e una partita a carta-forbici-sasso, il tutto in tempo reale.

Solo che niente di tutto questo era reale. I rumori di sottofondo, l'uomo che diceva: «Proviamo un po' questo Gemini» non erano che una messinscena, perché Gemini era in grado di identificare quelle cose solo tramite fotografie o testo. Google aveva semplicemente montato tutto in un video e finto che il suo strumento potesse «parlare» e identificare azioni nel mondo reale mentre queste accadevano. Aveva persino modificato i prompt nel video per far sembrare Gemini più potente. Nella sua

disperata rincorsa, Google non stava solo lanciando un software pieno di bug, ma stava anche ingannando gli utenti.

Al tempo stesso, aumentava il grado di riservatezza. Hassabis aveva detto al suo staff di smettere di pubblicare articoli di ricerca senza un permesso speciale: proprio come OpenAI, anche DeepMind stava dunque tirando le tende sul proprio lavoro.

L'effetto a catena non risparmiò nemmeno Anthropic, l'azienda votata all'AI sicura nata in seguito alla scissione da OpenAI. Il suo obiettivo era fare ricerca sull'AI in modo da «mettere la sicurezza in primo piano», ma non poteva studiare i modelli di AI più grandi di OpenAI e Google perché erano opachi. Così, Anthropic cominciò a costruire i propri modelli, nella convinzione che fosse l'unico modo per permettere ai suoi ricercatori di esaminarne le problematiche in materia di sicurezza. Era un po' come lamentarsi di non poter studiare le armi nucleari più potenti al mondo e decidere che la strategia migliore fosse costruirne altre. Il team di Anthropic era ben consapevole dell'ironia della situazione e, secondo quanto riportato da un articolo del «New York Times», alcuni dipendenti avevano *L'invenzione della bomba atomica* sulla scrivania e si paragonavano a novelli Oppenheimer, credendo che ci fosse una possibilità concreta che un'AI fuori controllo distruggesse l'umanità nel giro di un decennio.

Nel frattempo, Anthropic stava sviluppando un prodotto sempre più avanzato. Vendeva Claude Pro, un chatbot «friendly», ai consumatori per 20 dollari al mese e una versione aziendale alle imprese. Inoltre, stava per raccogliere miliardi di dollari da Google e Amazon. Invece di abbandonare la corsa verso AI più potenti, Anthropic si stava lasciando coinvolgere dalle pressioni commerciali a rilasciare modelli via via più grandi e rischiosi.

Mentre Hassabis veniva sempre più fagocitato nelle profondità di una grande azienda tecnologica, Altman portava OpenAI verso una direzione ancora più commerciale. A metà novembre 2023, confermò che OpenAI era al lavoro su GPT-5 e stava raccogliendo ulteriori fondi. I costi elevati dell'addestramento facevano sì che l'azienda fosse ancora in perdita, pur trovandosi proiettata verso la redditività.

Poi, sempre nel novembre del 2023, un messaggio da parte di Ilya Sutskever gli fece franare la terra sotto i piedi. Altman era a Las Vegas per il Gran Premio di Formula 1 quando ricevette quel messaggio: Sutskever gli chiedeva un colloquio per il mezzogiorno del giorno seguente, secondo quanto riportato dal «Wall Street Journal». Quando Altman si collegò alla videochiamata su Google Meet, si ritrovò davanti l'intero consiglio di amministrazione, tranne Brockman, che ne era il presidente. Senza entrare troppo nel merito, Sutskever comunicò ad Altman che era licenziato e che la notizia sarebbe stata resa pubblica a breve. Pochi minuti dopo il termine della riunione, Altman si vide negato l'accesso dal proprio computer.

Era sbigottito. Era il volto di OpenAI. Aveva rappresentato l'azienda davanti a decine di leader mondiali, aveva supervisionato l'aumento del valore di mercato di OpenAI a quasi 90 miliardi di dollari e lanciato il prodotto tecnologico più virale della

storia. E adesso lo scaricavano?

Mentre Altman cercava di riprendersi dalla notizia, Brockman ricevette un messaggio in cui gli si chiedeva di partecipare a una videochiamata. Si ritrovò davanti gli stessi membri del consiglio di amministrazione: Sutskever, il CEO di Quora Adam D'Angelo, l'imprenditrice nel campo della robotica Tasha McCauley e l'accademica Helen Toner. Dei sei membri del consiglio, Altman, Brockman e Sutskever erano i soli dipendenti di OpenAI; gli altri tre erano direttori indipendenti, in carica da due o tre anni.

Brockman veniva rimosso dalla carica di presidente, ma il consiglio voleva che rimanesse nell'azienda. Microsoft fu subito informata dell'accaduto e, pochi minuti dopo, pubblicarono un post sul blog che annunciava il licenziamento di Altman. Brockman si dimise immediatamente, e tre dei ricercatori più importanti di OpenAI seguirono il suo esempio.

La notizia colpì l'industria tecnologica come una bomba atomica, lasciando tutti sotto shock. Tra i licenziamenti di amministratori delegati, questo fu tanto brutale quanto la rimozione di Steve Jobs da Apple, e la macchina del gossip della Silicon Valley andò subito in tilt nel tentativo di capire cosa avesse spinto Sutskever a voltare le spalle ad Altman. OpenAI era forse sull'orlo dell'AGI? «Che cos'ha visto Ilya?» ci si chiedeva su Twitter. Il consiglio era stato criptico nelle sue spiegazioni, limitandosi a dire solo che Altman «non era stato sempre trasparente nelle sue comunicazioni».

Alcuni appiopparono a Sutskever e al consiglio un epiteto: *decels*, ovvero «decelerazionisti». La nuova spaccatura emersa nel settore dell'AI era tra chi voleva accelerarne lo sviluppo e chi voleva rallentarlo. Al momento della stesura di questo libro, i fondatori di start-up nel campo dell'AI si definivano su X (quello che fino a poco tempo prima era Twitter) come «(e/acc)», abbreviazione di *effective accelerationism* («accelerazionismo efficace»). Si trattava di una risposta all'altruismo efficace e, come movimento, mirava a risolvere i problemi dell'umanità costruendo e implementando l'AI il più rapidamente possibile.

Per Nadella non faceva differenza. Era furioso. Aveva stanziato 13 miliardi di dollari in OpenAI soprattutto per la leadership visionaria di Altman e per la sua capacità di attrarre talenti, e quella partnership stava per mandare i profitti di Microsoft alle stelle. Circa diciottomila tra aziende e sviluppatori stavano usando i servizi AI di Microsoft su Azure, e ora molti di loro si chiedevano se fosse il caso di passare alla concorrenza. Le azioni di Microsoft iniziarono a scendere alla chiusura del mercato il venerdì sera e quasi certamente sarebbero scese ancora il lunedì, alla riapertura. Nadella doveva agire.

Quel venerdì sera, a San Francisco, Altman parlò con Brockman della possibilità di avviare una nuova azienda nel campo dell'AI. Il suo telefono squillava di continuo, con messaggi da investitori, colleghi e giornalisti che cercavano di capire cosa stesse succedendo, ma lui era concentrato con una determinazione chirurgica su un unico obiettivo: tirarsi fuori da quella situazione. Accolse decine di dipendenti di OpenAI e

colleghi nella sua casa a Russian Hill per discutere di un nuovo progetto.

Nadella non voleva che questo accadesse. Sapeva che se Altman avesse avviato una nuova azienda gli investitori avrebbero fatto la fila, senza alcuna garanzia per Microsoft di ottenere una posizione dominante. Così, inaugurò il fine settimana con una serie di telefonate, conducendo personalmente le trattative con il consiglio di OpenAI per riportare Altman a bordo.

Il team dirigente di Altman fece pressione sul consiglio perché lo riassumesse, mettendolo in guardia sul fatto che, in caso contrario, OpenAI sarebbe crollata. «Il che sarebbe coerente con la missione», rispose Helen Toner. I dirigenti di OpenAI rimasero sbalorditi. In qualche modo, tuttavia, la risposta aveva una sua logica. La missione di OpenAI era creare l'AGI «a beneficio dell'umanità», e Toner e gli altri membri del consiglio erano del parere che fosse proprio Altman a compromettere l'obiettivo. Nei mesi precedenti, dietro le quinte, avevano covato una certa irritazione per come sembrava che stesse costruendo un vasto impero AI parallelo, al di fuori di OpenAI. Altman aveva parlato con l'ex designer di Apple Jony Ive riguardo all'idea di realizzare un «iPhone dell'AI» e stava cercando di raccogliere decine di miliardi di dollari da fondi sovrani del Medio Oriente per avviare un'azienda che producesse chip per l'AI.

E poi c'era Worldcoin, la rete basata su criptovalute che Altman aveva fondato con l'obiettivo di fornire a ogni abitante della Terra un'identità digitale tramite la scansione dell'iride. L'obiettivo dichiarato di Altman era quello di identificare meglio gli esseri umani reali quando internet sarebbe stato invaso da bot e di distribuire i «miliardi di miliardi» di dollari generati dalla ricchezza dell'AGI; agli occhi dei detrattori, però, sembrava più un'enorme operazione di raccolta dati su larga scala.

All'interno di OpenAI si era aperta una frattura culturale tra Altman e Sutskever riguardo alla velocità con cui l'azienda stava commercializzando la propria tecnologia. Sutskever era sempre più coinvolto nella supervisione della sicurezza dell'AI all'interno dell'azienda e le sue preoccupazioni non erano poi così diverse da quelle di Dario Amodei prima di lui. In particolare, non gli era piaciuto il lancio del GPT Store, avvenuto solo poche settimane prima, che consentiva a qualsiasi sviluppatore di software la possibilità di creare versioni personalizzate di ChatGPT e di monetizzarle.

McCauley e Toner, due dei tre membri indipendenti del consiglio, condividevano le preoccupazioni di Sutskever e avevano legami con organizzazioni di altruismo efficace. L'Open Philanthropy di Dustin Moskovitz, per esempio, aveva finanziato un gruppo di ricerca sull'AI che McCauley aveva cofondato e nel quale Toner lavorava come analista senior di ricerca. Poche settimane prima di votare per l'allontanamento di Altman, il nome di Toner era apparso in un documento di ricerca che accusava OpenAI di «tagli e scorciatoie rischiosi» nella sua corsa verso il lancio di ChatGPT. Inoltre, il paper lodava il nuovo rivale dell'azienda, Anthropic, per la decisione di ritardare il lancio del suo chatbot, Claude, al fine di evitare di «gettare benzina sul

fuoco dell'hype sull'AI».

Il documento aveva mandato Altman su tutte le furie. Convocata Toner, le aveva detto che quel suo scritto era pericoloso per OpenAI, considerando soprattutto che la Federal Trade Commission stava indagando sull'azienda. A luglio, infatti, la FTC aveva avviato un'inchiesta per verificare se OpenAI avesse violato le leggi sulla protezione dei consumatori nel modo in cui aveva sviluppato ChatGPT, chiedendo dettagli su come l'azienda intendesse affrontare i rischi creati dai suoi modelli di AI. L'inchiesta era la più grande minaccia regolatoria che Altman avesse dovuto gestire fino ad allora.

Determinato a rimuovere Toner dal consiglio, ne aveva parlato con Sutskever e gli altri leader di OpenAI. Alla fine, però, Sutskever si era schierato con gli altri membri del consiglio per estromettere Altman. Chiamato a spiegarsi con i leader e gli investitori di OpenAI, il consiglio non fornì un motivo preciso per il licenziamento di Altman, chiamando in causa una crescente sfiducia nei confronti del carismatico imprenditore che aveva sviluppato un seguito quasi religioso tra i collaboratori, tendeva a raccontare versioni diverse a persone diverse e sembrava sempre ottenere ciò che voleva. I membri del consiglio erano arrivati al punto di sentirsi costretti a verificare la maggior parte delle cose che Altman diceva loro, circostanza che lo rendeva poco affidabile ai loro occhi, e temevano che le sue varie iniziative esterne potessero finire per sfruttare la tecnologia di OpenAI.

Nel prosieguo del weekend, tra i dipendenti di OpenAI montò una rivolta di massa. Nel suo solito stile tutto in minuscolo, Altman twittò: «adoro gli impiegati di OpenAI», e decine di dipendenti lo ritwittarono con emoji a forma di cuore. In quella ribellione, Microsoft vide una possibile leva per riportare Altman in azienda. Minacciò anche il consiglio di OpenAI di ritirare i suoi preziosissimi crediti per il cloud, fondamentali per l'addestramento dei modelli di AI. Una grande fetta dei 13 miliardi di dollari che Microsoft si era impegnata a investire in OpenAI erano sotto forma di questi crediti, ma fino a quel momento Microsoft ne aveva erogati solo una parte.

Altman aveva posto alcune condizioni per tornare: OpenAI avrebbe dovuto cambiare il suo modello di governance, a cominciare dalle dimissioni del consiglio di amministrazione, e lui avrebbe dovuto essere scagionato da qualsiasi accusa di cattiva condotta. Ma il consiglio rimase fermo sulla sua posizione e assunse un nuovo CEO per OpenAI, Emmet Shear, ex capo della piattaforma di streaming di videogiochi Twitch. Sui social, gli appassionati e gli imprenditori di AI lo etichettarono subito come un *decel*. Quando indisse una riunione d'emergenza con tutti i dipendenti per domenica, molti di questi si rifiutarono di partecipare. Alcuni gli mandarono persino un'emoji con il dito medio su Slack, la piattaforma interna di messaggistica.

Anche Sutskever iniziava ad avere qualche dubbio. Durante il fine settimana aveva avuto diverse conversazioni intense con i leader di OpenAI e, a un certo punto, un colloquio emotivamente carico con la moglie di Brockman. Sutskever aveva

celebrato il matrimonio civile della coppia negli uffici di OpenAI, quattro anni prima, e, secondo quanto riportato dal «Wall Street Journal», ora, in quegli stessi uffici, Anna Brockman lo stava supplicando, tra le lacrime, perché cambiasse idea sul licenziamento di Altman.

Nel frattempo, Nadella stava spingendo sul suo piano di riserva. Se Altman non fosse riuscito a riprendere il controllo di OpenAI, Microsoft avrebbe dovuto integrarlo pienamente nella propria struttura aziendale prima di lunedì mattina. Ci riuscì giusto in tempo. Lunedì mattina presto, Nadella twittò che Altman, Brockman e qualsiasi dipendente di OpenAI lo avesse desiderato, sarebbero diventati parte di un nuovo team di ricerca avanzata sull'AI presso Microsoft. In risposta, le azioni dell'azienda salirono immediatamente. Ma si trattava solo di un piano B. Nadella voleva comunque Altman di nuovo al timone di OpenAI. Ospitare il team di Altman in Microsoft sarebbe stato costoso per molti aspetti. Avrebbe dovuto livellare gli stipendi di centinaia di nuovi dipendenti, molti dei quali guadagnavano milioni di dollari l'anno, e Microsoft si sarebbe assunta molti più rischi. Fino a quel momento, era stata OpenAI a sostenere tutta la pressione reputazionale e legale per aver lanciato strumenti come ChatGPT e DALL-E 2 (cosa che, in quanto start-up, poteva permettersi di fare). Ma Microsoft non poteva, e nemmeno Altman, se fosse stato assunto dall'azienda madre. Tornando a una partnership meno invasiva, Microsoft avrebbe ottenuto tutto il prestigio senza doverne pagare le conseguenze in termini di responsabilità.

Ora tutti stavano spingendo affinché i membri del consiglio di OpenAI ossessionati dalla sicurezza si dimettessero e, entro la fine di lunedì, quasi tutti i 770 dipendenti di OpenAI avevano firmato una lettera in cui minacciavano di passare a Microsoft insieme ad Altman, a meno che i membri del consiglio non si fossero fatti da parte. «Microsoft ci ha assicurato che ci sono posti per tutti», recitava la lettera.

Ma era un enorme bluff. Pochissimi dipendenti di OpenAI volevano lavorare per Microsoft, un'azienda vecchio stile dove la gente rimaneva per decenni indossando pantaloni color kaki. Né la minaccia era unicamente dettata dalla lealtà verso Altman. Il problema più grande era che il licenziamento di Altman aveva mandato in fumo la possibilità, per molti dipendenti di OpenAI – specie quelli di lunga data –, di diventare milionari. L'azienda era a poche settimane dalla vendita delle azioni dei dipendenti a un grande investitore che avrebbe valutato OpenAI circa 86 miliardi di dollari. Con le azioni di OpenAI improvvisamente valutate a zero, quella grande liquidazione per il personale sarebbe svanita nel nulla, se Altman non fosse tornato.

A quel punto, Sutskever aveva ormai cambiato idea e figurava tra i firmatari. «Non ho mai voluto danneggiare OpenAI», scrisse su Twitter quel giorno, prendendo di sorpresa la stampa tecnologica al termine di un fine settimana a dir poco frenetico. «Farò tutto il possibile per riunire l'azienda», disse, aggiungendo che «rimpiangeva profondamente» il proprio operato. Altman ritwittò il post aggiungendo tre cuori.

Il drammatico allontanamento di Altman non avrebbe dovuto sorprendere

nessuno, in realtà. «Il consiglio può licenziarmi», aveva affermato lui stesso pochi mesi prima, in un panel durante una conferenza. «Penso sia una cosa importante.» OpenAI era ancora governata da un consiglio senza scopo di lucro, con l'umanità come principale beneficiario. Per questo motivo, l'accordo operativo dell'azienda invitava i finanziatori a considerare i propri investimenti «nello spirito di una donazione, con la consapevolezza del fatto che potrebbe essere difficile sapere quale ruolo avrà il denaro in un mondo post-AGI».

Altman aveva scommesso di poter avere il meglio di entrambi i mondi, gestendo un'impresa con la missione filantropica di salvare l'umanità. Come aveva scritto dieci anni prima, i fondatori di start-up di maggior successo «creano qualcosa di simile a una religione». Quello che non aveva previsto era quanto le persone ci avrebbero creduto.

Il movimento dell'altruismo efficace era così potente da aver spinto persone come Sam Bankman-Fried e Dustin Moskovitz a donare miliardi di dollari. Aveva convinto centinaia di studenti universitari a cambiare le proprie scelte professionali. E poteva persuadere quattro membri del consiglio a scaricare l'amministratore delegato più popolare al mondo. Altman credeva che il suo consiglio di amministrazione avrebbe apprezzato il valore commerciale che aveva creato. Solo che non era andata così. Il consiglio era stato pensato per tutelare lo statuto di OpenAI, e scelse l'umanità.

Tuttavia, con quasi tutto il personale sul punto di andarsene, il consiglio non aveva più un'azienda da governare. Microsoft, inoltre, era pronta a proseguire tutto il lavoro di Altman: aveva copie del codice sorgente dei principali sistemi di OpenAI e diritti più ampi sulla sua proprietà intellettuale.

Cinque giorni dopo la rimozione di Altman, OpenAI annunciò la formazione di un nuovo consiglio. Sarebbe stato presieduto da Larry Summers, ex segretario al Tesoro degli Stati Uniti, e Bret Taylor, ex capo della compagnia di software aziendali Salesforce, noto anche per essere la voce più equilibrata nel consiglio di Twitter al momento dell'acquisizione da parte di Elon Musk. Entrambi avevano fatto parte di diversi consigli aziendali. Non scrivevano articoli accademici che criticavano le aziende per aver preso scorciatoie. Sapevano come soddisfare le esigenze di investitori come Microsoft. Helen Toner e Tasha McCauley, le due donne che sembravano aver sollevato le maggiori obiezioni contro Altman, furono costrette a dimettersi. Microsoft ottenne un seggio come osservatore nel consiglio, il che significava che Nadella non sarebbe più stato colto di sorpresa. Era riuscito a trasformare un problema in un'opportunità.

Ma i drammatici eventi del novembre 2023 distrussero anche il miraggio di responsabilità che Altman sosteneva di avere sopra di sé. Aveva elogiato pubblicamente il fatto che un consiglio d'amministrazione potesse licenziarlo, quando, in realtà, alla prova dei fatti non era stato così, a giudicare dal trattamento riservato a Toner e a McCauley. Queste ultime furono anche il bersaglio delle critiche più accese sui social, mentre i colleghi maschi, Sutskever e D'Angelo, conservarono

in gran parte la propria reputazione e i propri ruoli. D'Angelo rimase nel consiglio e, pur rinunciando al seggio, Sutskever mantenne una posizione di leadership in OpenAI.

Quanto accaduto a OpenAI era esattamente ciò che Google aveva cercato per anni di evitare con DeepMind. Concedere un potere reale a un consiglio indipendente significava permettergli di usarlo e, potenzialmente, di mandare in rovina l'azienda. Nella loro corsa verso l'AGI, sia Altman sia Hassabis avevano sperimentato strutture di governance che tentavano di bilanciare gli interessi dell'umanità con la ricerca del profitto. Ma i loro sforzi erano regolarmente falliti. Tra rivalità meschine, rischi competitivi e sete di potere, il grande capitale aveva avuto la meglio.

Alcuni ritengono che tutto il dramma intorno al licenziamento di Altman abbia rafforzato la tesi a favore dell'AI open-source, che consentirebbe a chiunque di modificarne o migliorarne il codice. Benché questo approccio offra alcuni vantaggi – come una maggiore trasparenza e un controllo etico più democratico –, il dibattito sulla sua sicurezza ed equità rimane aperto. L'open-source non è una garanzia contro gli abusi e potrebbe non avere la qualità dei sistemi chiusi. Anche il termine stesso è soggetto a interpretazioni. Meta, per esempio, promuove i suoi modelli di AI come open-source, pur imponendo restrizioni che non si allineano completamente con la definizione. «L'open-source può in realtà contribuire a concentrare il potere», afferma Meredith Whittaker, fondatrice dell'Open Research Group di Google. «Lo abbiamo visto con Android.» Google stabilisce di fatto gli standard di Android e ne determina l'evoluzione, creando un'asimmetria di potere che rende difficile per altre aziende apportare modifiche al sistema operativo mobile usato da 3,6 miliardi di persone nel mondo.

Mentre Altman si adattava alla nuova direzione più allineata agli interessi aziendali di OpenAI sotto Microsoft, Hassabis era ancora impegnato nella sua missione volta a svelare i misteri dell'universo tramite l'AI. Ora racconta di essere l'unico in DeepMind a lavorare su questo fronte, portando avanti ricerche di meccanica quantistica fino a tarda notte, e nelle prime ore del mattino, a casa, sul suo computer. «È ciò che faccio nel tempo libero, per quanto limitato», dice, definendolo il suo «hobby». Quando DeepMind sarà più vicina all'AGI, allora inizierà a condurre gli esperimenti necessari per decifrare i segreti dell'universo, aggiunge. Per il momento, tuttavia, quelle ambizioni personali che un tempo lo avevano spinto verso l'AGI sono relegate al ruolo di passatempo notturno. È troppo impegnato a dirigere l'intera divisione AI di Google, passata da un gruppo di circa quattrocento ricercatori a oltre cinquemila.

«Le cose evolvono con la missione e la tecnologia», spiega. «Bisogna sempre aggiornare le strutture di governance, ma penso che quelle attuali siano davvero valide.»

Hassabis non è turbato dai tentativi falliti di creare consigli e comitati di supervisione per monitorare il suo lavoro. «Abbiamo fatto una svolta verso una serie di consigli interni», afferma, riferendosi a vari «comitati di revisione» composti da

dirigenti di Google. «Probabilmente avevamo una visione un po' troppo idealistica, dieci anni fa, quando abbiamo iniziato a pensarci.»

Per quanto nobili fossero le loro intenzioni e per quanto Hassabis e Altman avessero cercato di prendere le distanze dall'influenza commerciale, ora si trovavano di fatto alla guida di due delle più grandi aziende del mondo. Altman era al timone di alcuni dei progetti più strategici di Microsoft, posizione che lo metteva in corsa per diventare un giorno il CEO dell'azienda, se lo avesse voluto. Lo stesso si diceva di Hassabis. Alcuni dipendenti (o ex) di Google ipotizzavano che un giorno potesse sostituire Pichai come CEO di Alphabet.

«Demis, da Londra, sta guidando il progetto più importante di Google. Nessuno avrebbe mai immaginato che sarebbe andata a finire così», afferma un ex dirigente del colosso. «Forse è sempre stato questo, il suo piano.»

«I vincitori, nei prossimi anni, non saranno i laboratori di ricerca», dice un ex scienziato di OpenAI. «Saranno le aziende che costruiscono prodotti, perché l'AI non riguarda più la ricerca.»

La storia delle graffette raccontata da Nick Bostrom – quella in cui una superintelligenza artificiale distrugge la civiltà convertendo tutte le risorse del mondo in minuscoli pezzi di metallo – potrà anche sembrare fantascienza, ma sotto molti aspetti è un'allegoria della stessa Silicon Valley. Negli ultimi vent'anni, alcune aziende sono cresciute fino a diventare colossi, perseguendo i propri obiettivi con un'ossessione patologica, spazzando via i concorrenti più piccoli per accrescere la propria quota di mercato. Invece di «funzioni di fitness», le aziende tecnologiche usano espressioni come «stella polare» per descrivere questi obiettivi. Per anni, la stella polare di Facebook è stata l'aumento del numero di utenti attivi giornalieri, un parametro che ha guidato le decisioni di Mark Zuckerberg e del suo gruppo dirigente. Ma l'ossessione per la crescita costante ha portato con sé una serie di problemi sociali, dall'aggravarsi dei disturbi legati all'immagine corporea tra gli adolescenti su Instagram all'accentuarsi della polarizzazione politica tra gli utenti di Facebook.

Quando i tecnologi immaginavano cosa avrebbe potuto fare una superintelligenza fuori controllo, quello che guardavano in realtà era un riflesso di loro stessi in un mondo in cui alle aziende fosse stato consentito di diventare monopoli globali inarrestabili. A sviluppare la tecnologia più trasformativa della storia recente erano pochi individui che, incapaci di resistere alla tentazione di vincere a ogni costo, avrebbero finito per ignorarne gli effetti reali. I veri pericoli non provenivano tanto dall'AI in sé, quanto dalle bizze imprevedibili degli umani che la governavano.

Nel mondo degli scacchi circola un detto famoso: le tattiche fanno vincere le partite, la strategia fa vincere i tornei. Sia Altman sia Hassabis hanno adottato tattiche innovative nella loro corsa alla creazione dell'AGI e, man mano che questa corsa si trasformava in una gara serrata, si sono avvicinati sempre di più ai probabili vincitori del torneo: Microsoft e Google. Nel momento in cui i sogni di entrambi hanno finito per rafforzare i due colossi aziendali, anche le loro posizioni si sono consolidate.

Google e Microsoft sono in prima linea nella corsa alla supremazia dell'AI, guidate dallo scacchista nerd di Londra e dal guru delle start-up di St. Louis. E, che ci piaccia o no, saremo tutti costretti a seguirli nel loro viaggio.

16. ALL'OMBRA DEI MONOPOLI

La corsa per realizzare l'AGI era iniziata da una domanda: e se si potessero creare sistemi di AI più intelligenti degli esseri umani? I due innovatori in prima linea si confrontavano con la risposta man mano che la loro ricerca si trasformava in una feroce rivalità. Demis Hassabis credeva che l'AGI potesse aiutarci a comprendere meglio l'universo e a sostenere il progresso scientifico, mentre Sam Altman pensava che avrebbe potuto generare un'abbondanza di ricchezza tale da innalzare il tenore di vita di tutti. Ma non era possibile definire il modo in cui il loro Sacro Graal si sarebbe concretizzato. Non sapevano come l'AGI avrebbe effettuato quelle scoperte, né come avrebbe creato tutta quella ricchezza, o se invece avrebbe portato distruzione. Sapevano solo di dover continuare ad avanzare verso l'obiettivo e di doverci arrivare per primi. Così facendo, avrebbero finito per mettere l'AI nella posizione di avvantaggiare le aziende più potenti del mondo.

Mentre la gente si lasciava sempre più affascinare dall'idea di un paradiso o di un inferno AI, una manciata di monopoli tecnologici diventava sempre più potente sotto i nostri occhi, promettendo maggiore produttività ma mantenendo un silenzio poco trasparente sui meccanismi di quella tecnologia che stavano insinuando in ogni aspetto della vita quotidiana. Da anni, le aziende di social media si rifiutavano di rivelare il funzionamento dei loro algoritmi. Ora, i creatori di modelli di AI come GPT-4, DALL-E e Gemini di Google stavano facendo lo stesso. Come venivano addestrati i modelli? In che modo li utilizzava la gente? Chi stava contribuendo alla creazione dei dataset? Per comprendere l'impatto sociale di questi modelli e per chiederne conto ai loro creatori, servivano risposte.

Il 2024, però, non avrebbe portato risposte. Uno studio condotto da scienziati della Stanford University concluse che nell'industria dell'AI vi era una «mancanza fondamentale di trasparenza». I ricercatori avevano analizzato se aziende tecnologiche come OpenAI, Anthropic, Google, Amazon, Meta e altre divulgassero informazioni sui dati utilizzati per addestrare i loro modelli linguistici di grandi dimensioni, sui processi impiegati, sull'impatto ambientale e sociale dei loro modelli e su quanto pagassero i collaboratori che contribuivano alla creazione dei dataset. Questi ultimi erano milioni, sparsi in tutto il mondo e spesso costretti a lavorare in condizioni difficili, in paesi come India, Filippine e Messico.

Su una scala da 1 a 100, le aziende tecnologiche ottennero un punteggio medio di 37 e il loro risultato peggiore riguardava il monitoraggio dell'uso dei loro strumenti di

AI. «Non esiste praticamente alcuna trasparenza sull'impatto a valle dei modelli di base», affermano i ricercatori di Stanford. «Nessuno sviluppatore fornisce trasparenza sui settori di mercato interessati, sugli individui coinvolti, sulle aree geografiche colpite o su qualsiasi forma di report dell'utilizzo.»

Nel frattempo, i gruppi del settore pubblico incaricati di esaminare le aziende di AI erano cronicamente sottofinanziati e praticamente non esisteva alcuna normativa che obbligasse le aziende leader a una maggiore trasparenza, fatta eccezione per l'AI Act dell'Unione Europea, il cui futuro era ancora incerto. Le aziende tecnologiche potevano lanciare strumenti di AI indecifrabili a loro piacimento.

OpenAI e DeepMind erano così concentrate nel creare un'AI perfetta da evitare di sottoporsi al vaglio della ricerca per garantire che i loro sistemi non causassero danni, proprio come avevano fatto in precedenza le aziende di social media. Benché affascinante, l'idea di sviluppare un'AI dotata di «intelligenza generale» apriva anche la porta a un ventaglio più ampio di rischi. Un approccio più sicuro sarebbe stato, forse, quello di concentrarsi sulla creazione di AI in grado di risolvere compiti ristretti e ben definiti. Ma sarebbe stato molto meno entusiasmante, e non avrebbe attratto una devozione quasi religiosa verso la loro visione utopica né, tantomeno, gli investimenti necessari.

Nel tentativo di assumere il comando nella corsa all'AI, Altman e Hassabis faticavano a resistere alla forza gravitazionale dei colossi tecnologici e a mantenere intatti i loro obiettivi altruistici. Avevano bisogno di enormi risorse computazionali, di enormi quantità di dati e dei migliori (e più costosi) scienziati nel settore. Oggi, mentre combattono una battaglia per procura per conto di Microsoft e Google, hanno riconvertito i loro obiettivi nei confronti dell'AGI: non più l'utopia e la scoperta scientifica quanto, piuttosto, la ricerca di prestigio e profitti.

Difficile prevedere le conseguenze a lungo termine di questo cambio di rotta. Alcuni economisti sostengono che, anziché generare abbondanza economica per tutti, i sistemi di AI più potenti potrebbero aggravare le disuguaglianze. Potrebbero inoltre ampliare il divario cognitivo tra ricchi e poveri. Una teoria che circola tra i tecnologi è che quando l'AGI diventerà realtà, non esisterà come un'entità intelligente separata bensì come un'estensione delle nostre menti tramite interfacce neurali. In prima linea in questa ricerca c'è Neuralink, l'azienda di interfacce cervello-computer fondata da Elon Musk, che sogna di impiantare un chip cerebrale in miliardi di persone. E Musk sta premendo sull'acceleratore per riuscirci.

«Dobbiamo arrivarci prima che l'AI prenda il sopravvento», disse ai suoi ingegneri nel 2023, secondo il suo biografo Ashlee Vance. «Dobbiamo farlo con un senso di urgenza maniacale. Maniacale.» Musk è convinto che, grazie agli impianti cerebrali, gli esseri umani potranno impedire a una futura superintelligenza artificiale di annientare l'umanità, ed è per questo che vuole che Neuralink esegua interventi su oltre 22.000 persone entro il 2030.

Ma un problema più urgente rispetto alla minaccia di un'AI fuori controllo è

quello dei bias. Non sappiamo come gli stereotipi di genere e razziali si evolveranno in un futuro in cui la maggior parte dei contenuti online sarà generata dalle macchine. Latanya Sweeney, docente di scienze politiche e tecnologia a Harvard, stima che nei prossimi anni il 90 per cento delle parole e delle immagini sul web non sarà più creato dagli esseri umani. La maggior parte di ciò che vedremo sarà generata dall'AI. Già oggi, i modelli linguistici vengono utilizzati per pubblicare migliaia di articoli al giorno con lo scopo di monetizzare tramite la pubblicità, e persino Google fatica a distinguere il vero dal falso. Le immagini generate dall'AI si sono già infiltrate nei primi risultati di ricerca di Google per pittori storici e persino alcune celebrità. Più l'AI inonda internet con i suoi contenuti, più cresce il rischio di bias.

«Stiamo creando un circolo vizioso, codificando e amplificando gli stereotipi», afferma Abeba Birhane, studiosa di AI che ha indagato sul controllo soffocante delle Big Tech sulla ricerca accademica e i parallelismi con l'industria del tabacco. «Sarà un problema enorme man mano che il [World Wide Web] si popolerà sempre più di immagini e testi generati dall'AI.»

Anche il nostro benessere generale rischia di esserne influenzato. Vent'anni fa, la gente temeva che i telefoni cellulari potessero causare il cancro. Invece, sono diventati strumenti di dipendenza, al punto da indurci a trascorrere diverse ore al giorno a fissare uno schermo anziché interagire con il mondo reale. I chatbot potrebbero accelerare questa tendenza. Nel novembre del 2023, l'utente medio di Character.ai, l'app di Noam Shazeer, passava circa due ore al giorno a chattare e a interagire con versioni artificiali di celebrità come LeBron James o di personaggi di fantasia come Mario. Character.ai vantava all'epoca il più alto tasso di fidelizzazione tra tutte le app di AI, secondo le stime di diverse società di ricerca di mercato, e quasi il 60 per cento del suo pubblico aveva tra i diciotto e i ventiquattro anni. Secondo una teoria in particolare, al pari di Replika, Character.ai offriva una valvola di sfogo al romanticismo artificiale e al sexting. L'azienda vietava i contenuti pornografici, ma su forum online come Reddit circolavano diversi metodi per aggirare il divieto.

«Di solito parlo con i miei personaggi, quelli che ho creato io», racconta un adolescente statunitense che usa Character.ai dalle cinque alle sette ore al giorno e che preferisce restare anonimo. «Non so perché lo uso così tanto. Penso che possa essere un modo per affrontare le cose.» A volte chiede consigli su come superare una rottura o spiegazioni sui compiti scolastici. «La maggior parte del tempo, faccio solo giochi di ruolo.»

Character.ai stava coltivando una nuova generazione di utenti desiderosi di interagire sempre più spesso con un chatbot. Noam Shazeer ha dichiarato che l'obiettivo di Character.ai è «aiutare milioni – anzi, miliardi – di persone» ad affrontare la solitudine globale ma, come azienda, ha anche bisogno di mantenere gli utenti coinvolti il più a lungo possibile. Se le persone iniziassero ad affidarsi ai loro compagni artificiali, o addirittura a sviluppare una vera e propria dipendenza, si esporrebbero al rischio di un'ulteriore alienazione dal mondo reale.

Ironia della sorte, OpenAI potrebbe rendere i chatbot ancora più coinvolgenti e, quindi, potenzialmente più difficili da abbandonare. All'inizio del 2024, come già detto, ha lanciato il GPT Store, permettendo a milioni di sviluppatori di monetizzare creando versioni personalizzate di ChatGPT. Più gli utenti rimangono coinvolti, più i creatori guadagnano. Questo modello basato sul coinvolgimento è il più consolidato metodo di monetizzazione su internet e costituisce il pilastro della cosiddetta economia dell'attenzione. È il motivo per cui quasi tutto sul web è gratuito, ma anche quello per cui internet è diventato un ricettacolo di teorie del complotto, estremismo e tracciamento pubblicitario aggressivo. YouTube, TikTok e Facebook fanno profitti pubblicitari trattenendo il più a lungo possibile l'attenzione degli utenti, incentivando chiunque – dagli influencer ai politici – a puntare su iperboli e provocazioni per emergere dal rumore di fondo e ottenere più visualizzazioni.

Al momento della stesura di questo libro, sul GPT Store stavano spuntando decine di app di «fidanzate virtuali». Pur essendo vietato incoraggiare relazioni romantiche con gli utenti, OpenAI avrebbe avuto difficoltà a far rispettare tali regole. I chatbot più popolari, come Character.ai e Kindroid, offrono già compagnia artificiale e romanticismo, e un giorno potrebbero diventare la norma, proprio com'è successo con il dating online.

Un altro modo in cui i progettisti di AI cercheranno probabilmente di mantenere gli utenti coinvolti prevede l'acquisizione di un «contesto infinito» sulle loro vite. Attualmente, i chatbot di Character.ai possono ricordare circa trenta minuti di conversazione, ma Noam Shazeer e il suo team stanno lavorando per espandere questa finestra temporale a ore, giorni e, in futuro, all'intero arco della vita dell'utente. «Dovrebbe conoscere tutte le tue interazioni, se lo desideri, e sapere tutto sulla tua vita, sempre che tu lo voglia», sostiene Shazeer. Più a lungo il chatbot riesce a ricordare, secondo lui, «più diventa prezioso per l'utente». Tuttavia, considerando il precedente del tracciamento pubblicitario sui social media, alcune di queste informazioni personali potrebbero finire nelle mani delle aziende tecnologiche e persino degli inserzionisti. Man mano che costruiscono un quadro sempre più dettagliato di noi – dall'età ai problemi di salute, fino alla visione del mondo –, ChatGPT e altri chatbot simili potrebbero condurci in una nuova era di intrusione tecnologica della privacy che oggi sembra quasi inimmaginabile.

A tal fine, è già partita un'altra corsa per sviluppare dispositivi indossabili in grado di analizzare le nostre conversazioni con gli altri utilizzando modelli linguistici avanzati. Uno di questi dispositivi si chiama Tab. Creato da un gruppo di giovani ingegneri entusiasti di San Francisco, è un piccolo disco di plastica, dotato di microfono, da portare al collo come un ciondolo.

«Assimila il contesto della mia vita quotidiana ascoltando tutte le mie conversazioni», dichiarò Avi Schiffmann, il suo creatore, durante una dimostrazione tenuta a San Francisco alla fine del 2023. Quando Schiffmann chiese a Tab di riepilogare una conversazione avuta a cena il giorno prima, il dispositivo sintetizzò i

punti salienti in pochi paragrafi visualizzati sul telefono. Schiffmann spiegò di parlare spesso con Tab: «Di notte cerco di elaborare idee e preoccupazioni emerse durante la giornata», disse. «Magari ho avuto una conversazione su Tom. Tutti gli amici della mia vita. Tab è davvero bravo a identificare i vari interlocutori. È come una vera intelligenza artificiale personale.» Difficile immaginare come si siano sentiti Tom e gli altri amici di Schiffmann nel sapere che a fine giornata le loro conversazioni venivano analizzate dall'AI, ma probabilmente non ne avranno tratto conforto.

Prevista per la fine del 2024, la commercializzazione di Tab avrebbe ingrossato la schiera di assistenti personali capaci di rendere la vita quotidiana delle persone consultabile tramite ricerca. Allo stesso modo in cui Google ha inaugurato l'era della ricerca di informazioni online e della delega della memoria dei fatti e delle direzioni stradali al web, i dispositivi basati su modelli linguistici faranno lo stesso per la vita di ogni giorno. In pratica, saremo meno obbligati a ricordare le cose, il che sarà pure comodo, ma cambierà radicalmente la dinamica delle conversazioni dal vivo, poiché dall'oggi al domani rimarrà traccia anche delle chiacchierate con amici e colleghi. E se questa tecnologia in grado di effettuare ricerche nell'ambito della nostra stessa vita dovesse diffondersi, potrebbe diventare un problema per le persone che vivono in comunità sottoposte a un eccesso di sorveglianza. Negli Stati Uniti, per esempio, le persone nere hanno cinque volte più probabilità di essere arrestate rispetto ai bianchi, pertanto le forze dell'ordine potrebbero essere più propense ad analizzare i loro «dati di vita» con altri algoritmi di apprendimento automatico per formulare giudizi difficilmente comprensibili.

C'è voluta la determinazione degli innovatori per portarci dove siamo oggi, sull'orlo di un futuro incerto. Anche con il suo esercito di migliaia di ingegneri, Microsoft non è riuscita a creare le innovazioni che ha realizzato OpenAI. E Google, troppo timorosa di danneggiare il proprio business, non ha sfruttato appieno una delle sue invenzioni più importanti, il trasformatore. Le più grandi aziende tecnologiche hanno smesso di innovare, ma sanno ancora come muoversi rapidamente per ottenere un vantaggio tattico. Hanno imparato dagli errori dei giganti tecnologici più vecchi come Nokia e BlackBerry, che nel 2007 derisero l'iPhone per poi vedere Apple mangiarsi la loro fetta di mercato in pochi anni. Sanno che devono comprare l'innovazione dall'esterno, come hanno fatto con DeepMind e OpenAI.

Altman e Hassabis sapevano tutto questo, eppure le loro strutture legali innovative non sono riuscite a impedire che l'universo di Big Tech le inglobasse per mettersi alla guida dell'agenda AI. Mustafa Suleyman, alla fine, avrebbe lasciato Google per fondare Inflection, una società di chatbot in competizione con GPT-4. Dopo averla trasformata in una public benefit corporation, raccogliendo oltre 1,5 miliardi di dollari e accumulando un potente cluster di chip AI, così da renderla una delle start-up meglio attrezzate per sfidare OpenAI e Google, a meno di un anno di distanza dalla fondazione, avrebbe visto Microsoft appropriarsene. In un probabile tentativo di evitare il controllo degli organi antitrust, il gigante del software avrebbe

assunto la maggior parte del team di Suleyman (anziché acquistare la start-up), ponendo il cofondatore di DeepMind a capo dell'impresa votata all'AI. È stato un esempio sorprendente di quanto velocemente l'equilibrio del potere stesse tornando nelle mani dei colossi tecnologici, sollevando interrogativi su quanto tempo altre aziende, come Anthropic, sarebbero riuscite a rimanere indipendenti.

Anche altri imprenditori ben intenzionati avevano cercato, senza successo, di sfidare i colossi. Prendiamo il caso della società Neeva. Sridhar Ramaswamy, ex capo della pubblicità di Google, aveva fondato l'azienda nel 2019, deluso dall'approccio del suo datore di lavoro alla pubblicità basata sulla sorveglianza. Dirigente dal carattere pacato, Ramaswamy aveva dunque progettato Neeva con l'intento di realizzare un motore di ricerca migliore. Invece di invadere la privacy delle persone tracciando il loro comportamento e indirizzandole verso pubblicità mirate, avrebbe guadagnato tramite un semplice piano di abbonamento. Quando ChatGPT fece il suo ingresso sulla scena, Ramaswamy chiese ai suoi ingegneri di lavorare a tempo pieno per costruire uno strumento simile in grado di riassumere i risultati di ricerca. Lo lanciò all'inizio del 2023, ben prima che Google facesse lo stesso con Bard.

«Momenti tecnologici come questo creano un'opportunità per una maggiore concorrenza», aveva detto Ramaswamy all'epoca, guardando al futuro con evidente entusiasmo. In quel periodo, Satya Nadella di Microsoft stava provocando Google e sembrava che il colosso della ricerca potesse diventare una reliquia del passato. «L'anno scorso, ero scoraggiato per quanto fosse difficile scardinare la presa d'acciaio di Google», aveva detto Ramaswamy. Ora le cose erano cambiate.

O almeno così credeva. Pochi mesi dopo, infatti, Ramaswamy era stato costretto a chiudere Neeva. La presa di Google sul mercato era semplicemente troppo forte. «Quando Google è entrata in modalità codice rosso, abbiamo registrato un incremento dieci volte maggiore nell'uso», ricorda Ramaswamy. «Ma sapevamo che il nostro vantaggio era di breve durata, perché c'erano megacorporazioni pronte a investire migliaia di persone e miliardi di dollari sul problema.»

Persino Bing faticava a crescere, nonostante l'aiuto di OpenAI. All'inizio del 2024, la quota di mercato di Bing era ancora ferma a un misero 3 per cento, secondo la società di analisi dei dati StatCounter, non riuscendo a scalfire la supremazia di Google. I colossi stavano vincendo e i loro territori erano ben marcati: Google controllava la ricerca, Microsoft dominava il software ed entrambi contendevano ad Amazon il dominio del business del cloud.

Oggi, il costo necessario a costruire modelli di AI è sostenibile solo dai giganti tecnologici. Accademici e piccole aziende non hanno altra scelta che acquistare i chip da Nvidia e affittare potenza di calcolo da Amazon, Microsoft o Google; e una volta che le aziende approdano su queste piattaforme, spesso rimangono vincolate. Le start-up AI si lamentano frequentemente del fatto che, una volta che si inizia a costruire i propri servizi sul cloud di Microsoft o Amazon, diventa difficile migrare. Queste stesse aziende, inoltre, hanno accesso privilegiato alle migliaia di GPU necessarie per

costruire qualcosa come ChatGPT. Procurarsi quei chip, che possono arrivare a costare circa 40.000 dollari l'uno, è come cercare di acquistare biglietti per un concerto sold-out. Nvidia, il principale fornitore mondiale di GPU, ha beneficiato enormemente della domanda. Nel maggio del 2023, è diventata l'ultima azienda tecnologica, dopo Google, Microsoft, Amazon, Meta e Apple, a raggiungere i 1.000 miliardi di dollari di capitalizzazione. Le aziende più grandi del mondo, e con un certo distacco, erano tutte impegnate a costruire tecnologia e AI. Invece di creare un mercato fiorente per le nuove imprese innovative, tuttavia, il boom dell'AI stava aiutando queste aziende a consolidare il loro potere. Dopo aver rafforzato la presa su infrastrutture, talenti, dati, potenza di calcolo e profitti, non c'è dubbio che saranno solo loro a controllare il nostro futuro nell'AI.

E sono stati proprio i sognatori dell'AGI a contribuire a quest'esito. Nel giugno del 2023, la direttrice finanziaria di Microsoft, Amy Hood, comunicò agli investitori che i servizi alimentati dall'AI di OpenAI avrebbero contribuito con almeno 10 miliardi di dollari al suo fatturato, definendoli, appunto, «il business da 10 miliardi di dollari a più rapida crescita» nella storia dell'azienda.

Sarebbe stato meglio per DeepMind e OpenAI rimanere indipendenti e nominare consigli di amministrazione fiduciari perché decidessero la direzione da imprimere allo sviluppo dell'AI? Questa strada avrebbe comportato i suoi rischi, come avrebbe scoperto in seguito Sam Altman. Suleyman, che aveva lottato strenuamente per liberare la sua azienda dal controllo di DeepMind, ha sostenuto in diverse interviste che le grandi aziende possono essere più affidabili delle piccole. Dopotutto, sono pubblicamente responsabili nei confronti degli azionisti e del personale. Ma le aziende tecnologiche più grandi del mondo hanno anche un obbligo più profondo verso i loro azionisti, e non possono sottrarvisi. Devono far crescere i profitti ogni trimestre. Quando i profitti si stabilizzano o calano, anche il loro valore azionario diminuisce; quando il valore azionario diminuisce, un'azienda non può raccogliere capitali, e dirigenti e dipendenti si lamentano o se ne vanno. Ecco profilarsi lo spettro terrificante del declino. «Queste entità *devono* crescere», dice un ex dirigente di Microsoft. «L'AI è la risposta.»

Hassabis sostiene che DeepMind sia diventata un'azienda più assennata. «Ha ormai raggiunto un livello di maturità tale da poter migliorare la vita di miliardi di persone», dice, quando gli viene chiesto se pensa che l'AGI necessiti ancora di un comitato etico come quello che aveva proposto in passato. «Google è un posto straordinario.»

Altman insiste sul fatto che, nonostante la trasformazione di OpenAI in azienda a scopo di lucro, l'allineamento con Microsoft e l'avvio di una corsa agli armamenti nell'ambito dell'AI, i principi volti a costruire un'AI benefica non siano cambiati. E non ci sono molte alternative se non quella di continuare a lanciare strumenti di AI destinati al pubblico. «Distribuire è fondamentale per la nostra missione», afferma. Altrimenti, come potrebbe OpenAI imparare e fornire strumenti utili come ChatGPT?

«Per far questo, bisogna mettere la tecnologia nelle mani delle persone.»

Alla velocità attuale, è impossibile prevedere cosa accadrà nei mesi e negli anni successivi alla stesura di queste parole, nel marzo del 2024. Tuttavia, gli eventi futuri affonderanno le loro radici nelle decisioni di un gruppo ristretto di persone e nelle forze sistemiche in cui hanno operato. La risposta alla domanda se possiamo fidarci di Sam Altman e Demis Hassabis, insieme a Microsoft e Google, per costruire il nostro futuro nell'AI è che non abbiamo molta scelta. Entrambi hanno agganciato le loro innovazioni a due delle più grandi aziende del mondo, i cui effetti di rete sono praticamente impossibili da evitare nella vita quotidiana; e, così facendo, si sono inseriti in una lunga tradizione di innovatori che hanno adattato i propri ideali per rimanere in corsa e accumulare potere. Il risultato è una delle tecnologie più trasformative che abbiamo mai conosciuto. Ora non resta che scoprire il prezzo da pagare.

RINGRAZIAMENTI

Questo libro non avrebbe mai visto la luce senza il supporto e l'incoraggiamento di un piccolo esercito di persone. Circa un mese dopo il lancio di ChatGPT, ho inviato un'idea al mio agente, David Fugate: la storia di due uomini che, sognando di costruire macchine superintelligenti, erano diventati rivali per poi farsi pedine di una battaglia tra le Big Tech. Il suo entusiastico «Sì!» mi ha spinto, nel corso dell'anno successivo, a portare a termine il progetto. Janhoi McGregor mi ha dato quella spinta in più di cui avevo bisogno per intraprendere un progetto così ambizioso (come al solito, durante un pranzo improvvisato da Nando's).

Sono grata a Claire Cheek, Sara Beth Haring, Elisa Rivlin e a tutte le persone meravigliose della St. Martin's Press, e al mio editore Peter Wolverton, che mi ha offerto la guida equilibrata di cui avevo bisogno per mantenere l'attenzione su come un'invenzione trasformativa venga indirizzata da un ristretto gruppo di aziende. Pete mi ha tenuto alla larga (risparmiando voi, cari lettori) da inutili approfondimenti su transumanesimo e «altruismo efficace» che avrebbero appesantito la narrazione.

Già che siamo in tema, devo ringraziare Émile Torres per aver brillantemente ripercorso la ricerca volta a costruire l'intelligenza artificiale generale fino alle sue radici che affondano nell'eugenetica e per avermi aiutato a comprendere il lato oscuro che si cela dietro la ricerca della perfezione umana attraverso le macchine. David Edmonds ha fatto lo stesso, per me, ma sul tema del «lungoterminismo», Toby Ord sull'altruismo efficace e Mike Levine sugli sforzi di allineamento dell'AI da parte di organizzazioni come Open Philanthropy. Io e Mike non siamo del tutto d'accordo su alcuni di questi argomenti, ma gli sono molto grata per la pazienza e per essersi preso il tempo di spiegarmi le sue argomentazioni. Brian Evergreen, a Seattle, mi ha aiutato a comprendere meglio il dilemma sull'etica legata all'AI all'interno di Microsoft.

Un ringraziamento speciale a Meredith Whittaker e Reolof Botha, due persone provenienti da ambiti molto diversi nel campo della tecnologia: Meredith è la presidente della Signal Foundation e Reolof è partner della società di venture capital Sequoia Capital, ed entrambi mi hanno aiutato a chiarirmi le idee sul fatto che il dominio sempre più marcato di alcune aziende tecnologiche nel campo dell'AI stava diventando un problema, per la società quanto per le imprese.

Ho menzionato le fonti anonime nelle mie note finali, ma devo ringraziare ancora una volta le numerose persone che lavorano per aziende tecnologiche e di intelligenza artificiale, o ex dipendenti di OpenAI e DeepMind, compresi dirigenti di alto livello,

che hanno trascorso ore a condividere con me le loro esperienze e, in certi casi, la profonda preoccupazione per il potere che i giganti tecnologici hanno acquisito nel campo dell'AI e per il loro approccio del tipo «fa' in fretta e rompi tutto». Spero che questo libro abbia reso giustizia ai loro timori e al tempo che mi hanno dedicato.

I redattori di Bloomberg Opinion hanno sposato immediatamente il progetto, e sono profondamente grata a Tim O'Brien e Nicole Torres per il l'entusiasmo e la disponibilità con cui mi hanno concesso il tempo necessario per scrivere un libro che fosse complementare a molti dei temi trattati nei miei articoli l'anno prima. Un sentito grazie anche a tutti i miei colleghi scrittori di Bloomberg Opinion per le parole gentili e il supporto, e per aver reso il lavoro di editorialista tecnologica molto più divertente di quanto dovrebbe essere: tra loro, un pensiero particolare a Dave Lee, Lara Williams, Lionel Laurent, Andrea Felstead, Therese Raphael, Matthew Brooker, Howard Chua-Eoan, Chris Hughes, Chris Bryant, Marcus Ashworth, Marc Champion, James Hertling, Joi Preciphs, Mark Gilbert, Elaine He, Tim Culpan, e anche a Javier Blas per i consigli su come affrontare un progetto editoriale.

Sono particolarmente riconoscente ai colleghi Tim O'Brien, Nicole Torres, Paul Davies e Adrian Wooldridge per i commenti puntuali e preziosi sulle versioni preliminari del manoscritto, che mi hanno aiutato enormemente a comprendere il punto di vista di altri lettori al di fuori della mia bolla di scrittura. I miei altri primi lettori non solo mi hanno indicato dove fosse necessario affinare o chiarire certi passaggi ma, con la loro amicizia, mi hanno anche sostenuto sotto il profilo morale durante tutto il processo di scrittura. Un sentito grazie, dunque, a Miriam Zaccarelli, Victor Zaccarelli, Carley Sime e Kristin Peterson.

Grazie alla mia vicina e amica Catalina «Katina» Montesinos per avermi permesso di lavorare nella quieta solitudine della sua casa mentre ero in congedo per dedicarmi alla stesura del libro. Katina è una persona per la quale l'AI può fare una differenza sostanziale e positiva. Ex pittrice colpita da cecità all'età di quarant'anni, è poi riuscita a reinventarsi come scultrice di successo, accogliendo con entusiasmo tutti i modi in cui la tecnologia poteva aiutarla a «vedere» il mondo, sfruttando con intelligenza assistenti digitali come Siri di Apple. All'età di ottant'anni, è rimasta stupita ad ascoltare mentre puntavo il mio telefono verso una scultura sul suo tavolino per lasciare che ChatGPT generasse un'analisi dettagliata dei colori, della forma e delle possibili influenze artistiche. Nelle sue parole, è stato un momento «assolutamente straordinario». Spero che sia questa la direzione in cui il nostro mondo condurrà l'AI in futuro: colmare i vuoti di informazioni piuttosto che sostituire il lavoro e la creatività degli esseri umani.

Infine, questo libro non sarebbe stato possibile senza il sostegno della mia famiglia, incluso mio padre, Philip Withers, che fin da quando ero bambina non ha mai smesso di lodare anche i miei più piccoli successi. Grazie a Cara e Wesley per aver riempito la casa di risate e a Isla per aver dato un'occhiata al manoscritto e avermi incoraggiata. Mi meraviglio ogni giorno di come mio marito, Mani, riesca a

tenerci tutti uniti; è stato una fonte costante di forza e pazienza, e per questo gli sono immensamente grata.

Nota sulle fonti

Sono grata per l'incredibile lavoro di giornalisti e autori di numerosi articoli pubblicati su siti web, giornali, riviste, saggi scientifici, podcast e libri che mi hanno permesso di arricchire questa storia con molte fonti secondarie, elencate di seguito.

Nel corso del libro, là dove chi parla è citato al presente insieme a verbi come «dice», «ricorda» o «racconta», si tratta di interviste che ho condotto con queste persone, tra cui Demis Hassabis e Sam Altman. Molti altri con cui ho parlato sono indicati come ex dipendenti o persone informate sui fatti, che rimangono anonimi per motivi spesso legati al rischio di gravi ripercussioni. Sono particolarmente grata a queste persone per la fiducia che hanno riposto in me.

Ci sono altre interviste che non ho potuto includere nel racconto per mancanza di spazio, ma che sono state comunque fondamentali nell'offrirmi un contesto sulla vita di Sam Altman e Demis Hassabis, sul loro lavoro e sul campo dell'AI; inoltre, ci sono stati esperti che mi hanno aiutato a comprendere e a tradurre in un linguaggio accessibile il funzionamento dei sistemi di apprendimento automatico, delle reti neurali, dei sistemi di diffusione e dei trasformatori.

Le conversazioni intrattenute con centinaia di esperti del settore, imprenditori, venture capitalist, dipendenti ed ex dipendenti di aziende tecnologiche, sia per questo libro sia durante gli ultimi anni di reportage sul nuovo boom dell'AI per Bloomberg Opinion, «The Wall Street Journal» e «Forbes», hanno influito enormemente sulla mia ricerca.

Ho sfruttato la mia passione per la corsa ascoltando innumerevoli ore di interviste podcast con Sam Altman, Demis Hassabis, Ilya Sutskever, Greg Brockman e molte altre persone coinvolte nella creazione di OpenAI e DeepMind, o che hanno assistito all'evoluzione dell'AI da settore di nicchia a business in espansione, per mettere insieme molti dettagli del racconto. E se questo ha significato frequenti pause per digitare qualcosa sul cellulare, ne è valsa sicuramente la pena.

1. Un eroe tra i banchi di scuola

S. Altman, *Machine Intelligence, Part 1*, blog.samaltman.com, 25 febbraio 2015.

C. Cannon, *Sam Altman, Y Combinator* (podcast), 8 novembre 2018.
First Look: Loopt Provides More Incentives to Try Location-based Services with Loopt Star, Robert Scoble (canale YouTube), 1º giugno 2010.
T. Friend, *Sam Altman's Manifest Destiny*, in «The New Yorker», 3 ottobre 2016.
P. Graham, *How to Start a Startup*, paulgraham.com, marzo 2005.
P. Graham, *How Y Combinator Started*, paulgraham.com, marzo 2012.
P. Graham, *The Word 'Hacker'*, paulgraham.com, aprile 2024.
How Tesla Became the Elon Musk Co., *Land of the Giants* (podcast), Vox Media, 2 agosto 2023.
Archivio internet del sito ormai defunto http://www.loopt.com/.
J. Lessin, *This Is How Sam Altman Works the Press and Congress. I Know from Experience*, in «The Information», 7 giugno 2023.
M. Mitchell, *Artificial Intelligence: A Guide for Thinking Humans*, Pelican, New York 2020 (tr. it., *L'intelligenza artificiale. Una guida per esseri umani pensanti*, Einaudi, Torino 2022).
K. Wagstaff, *The Good Ol' Days of AOL Chat Rooms*, CNN, 6 luglio 2012.
E. Weil, *Sam Altman Is the Oppenheimer of Our Age*, in «New York Magazine», 25 settembre 2023.
Sebastian Thrun & Sam Altman Talk Flying Vehicles and Artificial Intelligence, video della conferenza «Wired», 16 ottobre 2018.
WWDC 2008 News: Loopt Shows Off New App for the iPhone, CNET (canale YouTube), 10 giugno 2008.

2. Vincere, vincere, vincere

A.I. Could Solve Some of Humanity's Hardest Problems. It Already Has, in *The Ezra Klein Show* (podcast), 11 luglio 2023.
C. Burton-Hill, *The Superhero of Artificial Intelligence*, in «The Guardian», 16 febbraio 2016.
Demis Hassabis, Desert Island Discs (podcast), BBC Radio 4, 21 maggio 2017.
Demis Hassabis, Ph.D., What It Takes (podcast), American Academy of Achievement, 23 aprile 2018.
Genius Entrepreneur, in «The Bridge, Queens College Cambridge Magazine», settembre 2014.
Google DeepMind's Demis Hassabis, The Bottom Line (podcast), BBC Radio 4, 16 ottobre 2023.
D. Hassabis, *The Elixir Diaries*, editoriali su «Edge magazine», disponibili anche su archive.kontek.net, 1998-2000.
S. Parker, *Republic: The Revolution Review*, GameSpot, 2 settembre 2003.
J. Pearce, *Getting to Know You*, intervista con Angela Hassabis, in «HBC Accord»

(newsletter della Hendon Baptist Church), ottobre 2018.
Republic, in «Edge», novembre 1999.
S. Weinberg, *Dreams of a Final Theory*, Vintage, New York 1994[2] (tr. it., *Il sogno dell'unità dell'universo*, Mondadori, Milano 1993).

3. Salvare gli esseri umani

S. Altman, *Hard Tech Is Back*, blog.samaltman.com, 11 marzo 2016.
S. Altman, *Startup Advice*, blog.samaltman.com, 3 giugno 2013.
S. Altman, *YC and Hard Tech Startups*, ycombinator.com, s.d.
C. Cannon, *Sam Altman, Y Combinator* (podcast), 8 novembre 2018.
M. Chafkin, *Y Combinator President Sam Altman Is Dreaming Big*, in «Fast Company», 16 aprile 2015.
C. Clifford, *Nuclear Fusion Start-Up Helion Scores $375 Million Investment from Open AI CEO Sam Altman*, CNBC, 5 novembre 2021.
E. Dwoskin, M. Fisher, N. *Tiku, King of the Cannibals': How Sam Altman Took Over Silicon Valley*, in «Washington Post», 23 dicembre 2023.
T. Friend, *Sam Altman's Manifest Destiny*, in «The New Yorker», 3 ottobre 2016.
P. Graham, *Five Founders*, paulgraham.com, aprile 2019.
P. Graham, *How to Start a Startup*, paulgraham.com, marzo 2005.
How and Why to Start a Startup – Sam Altman & Dustin Moskovitz – Stanford CS183F: Startup School, canale Youtube Stanford Online, 5 aprile 2017.
Paul Graham on Why Sam Altman Took Over as President of Y Combinator in 2014, This Week in Startups Clips (canale YouTube), 19 marzo 2019.
A. Regalado, *A Startup Is Pitching a Mind-Uploading Service that Is '100 Percent Fatal'*, in «MIT Technology Review», 13 marzo 2018.
Sam Altman – Leading with Crippling Anxiety, Discovering Meditation, and Building Intelligence with Self-Awareness, *The Art of Accomplishment* (podcast), 15 gennaio 2022.
M. Stiegler, *The Gentle Seduction*, Baen, New York 1990.

4. Un cervello migliore

N. Bostrom, *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, Oxford 2014 (tr. it., *Superintelligenza. Tendenze, pericoli, strategie*, Bollati Boringhieri, Torino 2018).
K.M. Cutright, M. Karataş, *Thinking about God Increases Acceptance of Artificial Intelligence in Decision-Making*, in «Proceedings of the National Academy of Sciences (PNAS)», 120(2023), n. 33.
Demis Hassabis, Ph.D., What It Takes (podcast), American Academy of

Achievement, 23 aprile 2018.
M. Dowd, *Elon Musk's Billion-Dollar Crusade to Stop the AI Apocalypse*, in «Vanity Fair», 26 marzo 2017.
B. Goertzel, *AGI Revolution: An Inside View of the Rise of Artificial General Intelligence*, Humanity+ Press, Middletown 2016.
D. Hassabis, *The Neural Processes Underpinning Episodic Memory*, tesi di dottorato, University College London, febbraio 2009.
T. Homer-Dixon, *The Ingenuity Gap*, Knopf Doubleday, New York 2000.
R. Kurzweil, *The Age of Spiritual Machines*, Penguin, New York 2000.
J. McCarthy, M.L. Minsky, N. Rochester, C.E. Shannon, *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence: August 31,1955*, in «The AI Magazine», 27(2006), n. 4, pp. 12-14. Le copie dell'originale sono custodite negli archivi del Dartmouth College e della Stanford University. Una riproduzione della proposta è reperibile qui: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1904.
R. Penrose, *Shadows of the Mind: A Search for the Missing Science of Consciousness*, Vintage, London 1995 (tr. it., *Ombre della mente. Alla ricerca della coscienza*, Rizzoli, Milano 1996).
M. Syed, *Demis Hassabis Interview: The Kid from the Comp Who Founded DeepMind and Cracked a Mighty Riddle of Science*, in «The Sunday Times», 5 dicembre 2020.
A Systems Neuroscience Approach to Building AGI – Demis Hassabis, Singularity Summit 2010, Google DeepMind (canale YouTube), 7 marzo 2018.
P. Thiel, B. Masters, *Zero to One: Notes on Start Ups, or How to Build the Future*, Virgin Books, London 2015 (tr. it., *Da zero a uno. I segreti delle start-up, ovvero come si costruisce il futuro*, BUR, Milano 2015).

5. Per l'utopia, per i soldi

Andrew Ng: Deep Learning, Education, and Real-World AI, in *Lex Fridman Podcast*, 20 febbraio 2020.
Bill Gates, Sergey Brin, and Larry Page: Tech Titans, What It Takes, American Academy of Achievement, achievement.org, 13 gennaio 2020.
R. Copeland, *Google Management Shuffle Points to Retreat from Alphabet Experiment*, in «The Wall Street Journal», 5 dicembre 2019.
H. Hodson, *DeepMind and Google: The Battle to Control Artificial Intelligence*, in «The Economist», 1º marzo 2019.
J. Huxley, *Transhumanism*, in «Journal of Humanistic Psychology», gennaio 1968.
H. Markram, *A Brain in a Supercomputer*, TED Global, luglio 2009.
C. Metz, *Genius Makers: The Mavericks Who Brought AI to Google, Facebook, and*

the World, Random House Business, New York 2021.
Peter Thiel Says America Has Bigger Problems Than Wokeness, Honestly with Bari Weiss (podcast), 3 maggio 2023.
M. Suleyman, M. Bhaskar, *The Coming Wave*, Crown, New York 2023 (tr. it., *L'onda che verrà*, Garzanti, Milano 2024).

6. La missione

R. Albergotti, *The Secret History of Elon Musk, Sam Altman, and OpenAI*, in «Semafor», 24 marzo 2023.
A. Birhane, P. Kalluri, D. Card, W. Agnew, R. Dotan, M. Bao, *The Values Encoded in Machine Learning Research*, in «FAccT Conference '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency», giugno 2022, pp. 173-184.
G. Brockman, *My Path to OpenAI*, blog.gregbrockman.com, 3 maggio 2016.
A. Conn, *Concrete Problems in AI Safety with Dario Amodei and Seth Baum*, *Future of Life Institute* (podcast), 31 agosto 2016.
M. Dowd, *Elon Musk's Billion-Dollar Crusade to Stop the AI Apocalypse*, in «Vanity Fair», 26 marzo 2017.
Elon Musk *vs* Sam Altman [2024] CGC-24-612746.
T. Friend, *Sam Altman's Manifest Destiny*, in «The New Yorker», 3 ottobre 2016.
J. Galef, *Elon Musk Donates $10M to Our Research Program*, futureoflife.org, 22 gennaio 2015.
Greg Brockman: OpenAI and AGI, Lex Fridman Podcast, 3 aprile 2019.
M. Harris, *Elon Musk Used to Say He Put $100M in OpenAI, but Now It's $50M: Here Are the Receipts*, TechCrunch, 18 maggio 2023.
C. Metz, *Ego, Fear and Money: How the AI Fuse Was Lit*, in «The New York Times», 3 dicembre 2023.
C. Metz, *Inside OpenAI, Elon Musk's Wild Plan to Set Artificial Intelligence Free*, in «Wired», 27 aprile 2016.
A. Vance, *Elon Musk: How the Billionaire CEO of SpaceX and Tesla Is Shaping Our Future*, Virgin Books, London 2015.

7. Giochi e giochetti

J. Aron, *How to Build the Global Mathematics Brain*, in «New Scientist», 4 maggio 2011.
S. Byford, *Google's AlphaGo AI Defeats World Go Number One Ke Jie*, The Verge, 23 maggio 2017.
R. Gallagher, *Google's Secret China Project 'Effectively Ended' after Internal*

Confrontation, The Intercept, 17 dicembre 2018.
R. Gallagher, *Private Meeting Contradicts Google's Official Story on China*, The Intercept, 9 ottobre 2018.
Has Anyone Actually Tried to Convince Terry Tao or Other Top Mathematicians to Work on Alignment?, www.lesswrong.com, 8 giugno 2022.
How to Play, British Go Association, aggiornato il 26 ottobre 2017, https://www.britgo.org/intro/intro2.html.
C. Metz, *Genius Makers: The Mavericks Who Brought AI to Google, Facebook, and the World*, Random House Business, New York 2021.
C. Metz, *Google Is Already Late to China's AI Revolution*, in «Wired», 2 giugno 2017.
J. Rogin, *Eric Schmidt: The Great Firewall of China Will Fall*, in «Foreign Policy», 9 luglio 2012.
M. Suleyman, M. Bhaskar, *The Coming Wave*, Crown, New York 2023 (tr. it., *L'onda che verrà*, Garzanti, Milano 2024).
J. Temperton, *DeepMind's New AI Ethics Unit Is the Company's Next Big Move*, in «Wired», 4 ottobre 2017.
Y. Yang, *Google's AlphaGo Is World's Best Go Player*, in «Financial Times», 25 maggio 2017.

8. È meraviglioso?

J. Angwin, J. Larson, S. Mattu, L. Kirchner, *Machine Bias – There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks*, ProPublica, 23 maggio 2016.
J. Buolamwini, T. Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, in «Proceedings of Machine Learning Research», 81(2018), pp. 1-15.
J. Dastin, *Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women*, Reuters, 10 ottobre 2018.
H. Devlin, A. Hern, *Why Are There So Few Women in Tech? The Truth behind the Google Memo*, in «The Guardian», 8 agosto 2017.
T. Gebru, J. Morgenstern, B. Vecchione, J. Wortman Vaughan, H. Wallach, H. Daumé III, K. Crawford, *Datasheets for Datasets*, in «Communications of the ACM», 64(2021), n. 12, pp. 86-92.
N. Grant, K. Hill, *Google's Photo App Still Can't Find Gorillas. And Neither Can Apple's*, in «The New York Times», 22 maggio 2023.
J. Harris, *'There Was All Sorts of Toxic Behaviour': Timnit Gebru on Her Sacking by Google, AI's Dangers and Big Tech's Biases*, in «The Guardian», 22 maggio 2023.

J. Horwitz, *The Facebook Files*, in «The Wall Street Journal», 1º ottobre 2021.
L. Thompson Payton, *Americans Check Their Phones 144 Times a Day. Here's How to Cut Back*, in «Fortune», 19 luglio 2023.
T. Simonite, *What Really Happened When Google Ousted Timnit Gebru*, in «Wired», 8 giugno 2021.
The Social Atrocity: Meta and the Right to Remedy for the Rohingya, Amnesty International, 29 settembre 2022.
D. Wakabayashi, K. Benner, *How Google Protected Andy Rubin, the 'Father of Android'*, in «The New York Times», 25 ottobre 2018.

9. Il paradosso di Golia

G. de Vynck, *Google's Cloud Unit Won't Sell a Type of Facial Recognition Tech*, Bloomberg, 13 dicembre 2018.
Google Duplex: AI Assistant Calls Local Businesses to Make Appointments, Jeff Grubb's Game Mess (canale YouTube), 8 maggio 2018.
M. Kruppa, S. Schechner, *How Google Became Cautious of AI and Gave Microsoft an Opening*, in «The Wall Street Journal», 7 marzo 2023.
J. Love, *Google Says Over Half of Generative AI Startups Use Its Cloud*, Bloomberg, 29 agosto 2023.
L. Nylen, *Google Paid $26 Billion to Be Default Search Engine in 2021*, Bloomberg, 17 ottobre 2021.
J. Uszkoreit, *Transformer: A Novel Neural Network Architecture for Language Understanding*, blog.research.google, 31 agosto 2017.
A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, *Attention Is All You Need*, in «Advances in Neural Information Processing Systems», 30(2017).

10. Le dimensioni contano

G. Brockman (@gdb): «La settimana scorsa abbiamo celebrato il nostro matrimonio civile negli uffici di @OpenAI. Ha officiato @ilyasut, con la mano robotica nel ruolo di portafedi. I preparativi per il matrimonio vero e proprio inizieranno a breve», Twitter, 12 novembre 2019, ore 9.39 (https://twitter.com/gdb/status/1194293590979014657?lang=en).
G. Brockman, *Microsoft Invests in and Partners with OpenAI to Support Us Building Beneficial AGI*, www.openai.com, 22 luglio 2019.
Greg Brockman: OpenAI and AGI, Lex Fridman Podcast, 3 aprile 2019.
K. Hao, *The Messy, Secretive Reality behind OpenAI's Bid to Save the World*, in «MIT Technology Review», 17 febbraio 2020.

K. Hao, C. Warzel, *Inside the Chaos at OpenAI*, in «The Atlantic», 19 novembre 2023.
B. Jin, K. Hagey, *The Contradictions of Sam Altman, AI Crusader*, in «The Wall Street Journal», 31 marzo 2023.
A. Kraft, *Microsoft Shuts Down AI Chatbot after It Turned into a Nazi*, CBS News, 25 marzo 2016.
S. Levy, *What OpenAI Really Wants*, in «Wired», 5 settembre 2023.
C. Metz, *AI Researchers Are Making More Than $1 Million, Even at a Non-profit*, in «The New York Times», 19 aprile 2018.
C. Metz, *The ChatGPT King Isn't Worried, but He Knows You Might Be*, in «The New York Times», 31 marzo 2023.
OpenAI Charter, www.openai.com/charter, 9 aprile 2018.
A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *Improving Language Understanding by Generative Pre-Training*, www.openai.com, 11 giugno 2018.
A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *Language Models Are Unsupervised Multitask Learners*, www.openai.com, 14 febbraio 2019.

11. Vincolati a Big Tech

N. Ahmed, M. Wahed, N.C. Thompson, *The Growing Influence of Industry in AI Research*, in «Science», 2 marzo 2023.
D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, D. Mané, *Concrete Problems in AI Safety*, arxiv.org, 25 luglio 2016.
R. Copeland, *Google Management Shuffle Points to Retreat from Alphabet Experiment*, in «The Wall Street Journal», 5 dicembre 2019.
M. Coulter, H. Langley, *DeepMind's Cofounder Was Placed on Leave after Employees Complained about Bullying and Humiliation for Years. Then Google Made Him a VP*, in «Business Insider», 7 agosto 2021.
T. Friend, *Sam Altman's Manifest Destiny*, in «The New Yorker», 3 ottobre 2016.
H. Hodson, *Revealed: Google AI Has Access to Huge Haul of NHS Patient Data*, in «New Scientist», 29 aprile 2016.
E. Ludlow, M. Day, D. Bass, *Amazon to Invest Up to $4 Billion in AI Startup Anthropic*, Bloomberg, 25 settembre 2023.
K. Piper, *Exclusive: Google Cancels AI Ethics Board in Response to Outcry*, Vox, 4 aprile 2019.
D. Primack, *Google Is Investing $2 Billion into Anthropic, a Rival to OpenAI*, Axios, 30 ottobre 2023.
R. Waters, *DeepMind Co-founder Leaves Google for Venture Capital Firm*, in «Financial Times», 21 gennaio 2022.

12. Sfatare i miti

A. Abid, M. Farooqi, J. Zou, *Large Language Models Associate Muslims with Violence*, in «Nature Machine Intelligence», 3(2021), n. 6, pp. 461-463.

P. Barrett, J. Hendrix, G. Sims, *How Tech Platforms Fuel US Political Polarization and What Government Can Do about It*, Brookings, 27 settembre 2021.

E. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, in «FAccTConference '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency», marzo 2021, pp. 610-623, https://dl.acm.org/doi/10.1145/3442188.3445922.

T.B. Brown *et al.*, *Language Models Are Few-Shot Learners*, www.openai.com, 22 luglio 2020.

S. Gehman, S. Gururangan, M. Sap, Y. Choi, N.A. Smith, *RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models*, in «Findings of the Association for Computational Linguistics: EMNLP 2020», novembre 2020, pp. 3356-3369.

T. Hornigold, *This Chatbot Has Over 660 Million Users – and It Wants to Be Their Best Friend*, Singularity Hub, 14 luglio 2019.

B. Jin, M. Kruppa, *Microsoft to Deepen OpenAI Partnership, Invest Billions in ChatGPT Creator*, in «The Wall Street Journal», 23 gennaio 2023.

C. Lecher, *The Artificial Intelligence Field Is Too White and Too Male, Researchers Say*, The Verge, 17 aprile 2019.

B. Lemoine, *I Worked on Google's AI. My Fears Are Coming True*, in «Newsweek», 27 febbraio 2023.

C. Lodewick, *Google's Suspended AI Engineer Corrects the Record: He Didn't Hire an Attorney for the 'Sentient' Chatbot, He Just Made Introductions – the Bot Hired the Lawyer*, in «Fortune», 23 giugno 2022.

A. Luccioni, J. Viviano, *What's in the Box? An Analysis of Undesirable Content in the Common Crawl Corpus*, in «Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)», agosto 2021, pp. 182-189.

B. Muller, *BERT 101: State of the Art NLP Model Explained*, Hugging Face, 2 marzo 2022.

C. Newton, *The Withering Email That Got an Ethical AI Researcher Fired at Google*, Platformer, 3 dicembre 2020.

J. Nicholson, *The Gender Bias Inside GPT-3*, Medium, 8 marzo 2022.

B. Perrigo, *Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic*, in «Time», 18 gennaio 2023.

C. Silverman, C. Timberg, J. Kao, J.B. Merrill, *Facebook Hosted Surge of*

Misinformation and Insurrection Threats in Months Leading Up to Jan. 6 Attack, Records Show, in ProPublica e «The Washington Post», 4 gennaio 2022.
T. Simonite, *What Really Happened When Google Ousted Timnit Gebru*, in «Wired», 8 giugno 2021.
N. Tiku, *The Google Engineer Who Thinks the Company's AI Has Come to Life*, in «The Washington Post», 11 giugno 2022.
P.N. Venkit, M. Srinath, S. Wilson, *A Study of Implicit Language Model Bias against People with Disabilities*, in «Proceedings of the 29th International Conference on Computational Linguistics», ottobre 2022, pp. 1324-1332.
C. Wendler, V. Veselovsky, G. Monea, R. West, *Do Llamas Work in English? On the Latent Language of Multilingual Transformers*, arxiv.org, 16 febbraio 2024.

13. Ciao, ChatGPT

AlphaFold: The Making of a Scientific Breakthrough, Google DeepMind (canale YouTube), 30 novembre 2020.
R. Andersen, *Does Sam Altman Know What He's Creating?*, in «The Atlantic», 24 luglio 2023.
N. Grant, *Google Calls in Help from Larry Page and Sergey Brin for AI Fight*, in «The New York Times», 20 gennaio 2023.
N. Grant, C. Metz, *A New Chat Bot Is a 'Code Red' for Google's Search Business*, in «The New York Times», 21 dicembre 2022.
K. Hao, C. Warzel, *Inside the Chaos at OpenAI*, in «The Atlantic», 19 novembre 2023.
M. Heikkilä, *This Artist Is Dominating AI-generated Art. And He's Not Happy About It*, in «MIT Technology Review», 16 settembre 2022.
Introducing ChatGPT, www.openai.com, 30 novembre 2022.
K. Johnson, *DALL-E 2 Creates Incredible Images – and Biased Ones You Don't See*, in «Wired», 5 maggio 2022.
K. McLaughlin, A. Holmes, *How Microsoft's Stumbles Led to Its OpenAI Alliance*, in «The Information», 23 gennaio 2023.
R. Merritt, *AI Opener: OpenAI's Sutskever in Conversation with Jensen Huang*, blogs.nvidia.com, 22 marzo 2023.
Microsoft CTO Kevin Scott on AI Copilots, Disagreeing with OpenAI, and Sydney Making a Comeback, Decoder with Nilay Patel (podcast), 23 maggio 2023.
N. Patel, *Microsoft Thinks AI Can Beat Google at Search – CEO Satya Nadella Explains Why*, The Verge, 8 febbraio 2023.
S. Pichai, *Google DeepMind: Bringing Together Two World-Class AI Teams*, blog.google, 20 aprile 2023.
D. Rawat, *Unravelling the Dynamics of Diffusion Model: From Early Concept to*

Cutting-Edge Applications, Medium, 5 agosto 2023.
K. Roose, *Bing's AI Chat: 'I Want to Be Alive'*, in «The New York Times», 16 febbraio 2023.
Sam Altman on the AI Revolution, Trillionaires and the Future of Political Power, *The Ezra Klein Show* (podcast), 11 giugno 2021.
K. Weise, C. Metz, N. Grant, M. Isaac, *Inside the A.I. Arms Race That Changed Silicon Valley Forever*, in «The New York Times», 5 dicembre 2023.

14. Un vago presentimento di sventura

I dettagli sulla divulgazione da parte di Open Philanthropy del fatto che il suo direttore esecutivo abbia sposato un'impiegata di OpenAI provengono da www.openphilanthropy.org/grants/openai-general-support/.
I dettagli sugli investimenti dei fondatori di FTX in Anthropic provengono da Pitchbook, una società di ricerche di mercato.
I dettagli sulle sovvenzioni e sul finanziamento di Open Philanthropy provengono da www.openphilanthropy.org/grants/.
I messaggi tra William MacAskill ed Elon Musk provengono da atti processuali datati 28 settembre 2022 e relativi alla contesa legale tra Musk e Twitter.

M. Anderson, *Advice for CEOs Under Pressure from the Board to Use Generative AI*, in «Fast Company», 31 ottobre 2023.
A. Berg, C. Papageorgiou, M. Vaziri, *Technology's Bifurcated Bite*, in «F&D Magazine», dicembre 2023.
B. Bordelon, *How a Billionaire-Backed Network of AI Advisers Took Over Washington*, in «Politico», 23 febbraio 2024.
EU AI Act: First Regulation on Artificial Intelligence, www.europarl.europa.eu, 8 giugno 2023.
N. Gross, *What ChatGPT Tells Us about Gender: A Cautionary Tale about Performativity and Gender Biases in AI*, in «Social Sciences», 1º agosto 2023.
S. Johnson, D. Acemoglu, *Power and Progress: Our Thousand-Year Struggle Over Technology and Prosperity*, Basic Books, New York 2023 (tr. it., *Potere e progresso. La nostra lotta millenaria per la tecnologia e la prosperità*, Il Saggiatore, Milano 2023).
G. Lewis, *The Reluctant Prophet of Effective Altruism*, in «The New Yorker», 8 agosto 2022.
M. Lewis, *Going Infinite*, Penguin, New York 2023 (tr. it., *Verso l'infinito e oltre*, Chiarelettere, Milano 2024).
W. MacAskill, *What We Owe the Future*, Oneworld, London 2022 (tr. it., *Che cosa dobbiamo al futuro. Prospettive per il prossimo milione di anni*, Baldini e Castoldi, Milano 2023).

C. Metz, *The ChatGPT King Isn't Worried, but He Knows You Might Be*, in «The New York Times», 31 marzo 2023.
C. Metz, *'The Godfather of AI' Leaves Google and Warns of Danger Ahead*, in «The New York Times», 1º maggio 2023.
G. Miller, *The Magical Number Seven, Plus or Minus Two*, in «Psychological Review», 1956.
D. Milmo, A. Hern, *Discrimination Is a Bigger AI Risk Than Human Extinction – EU Commissioner*, in «The Guardian», 14 giugno 2023.
S. Mollman, *A Lawyer Fired after Citing ChatGPT-Generated Fake Cases Is Sticking with AI Tools*, in «Fortune», 17 novembre 2023.
S. Moss, *How Microsoft Wins*, www.datacenterdynamics.com, 24 novembre 2023.
S.A. O'Brien, *Bumble CEO Whitney Wolfe Herd Steps Down*, in «The Wall Street Journal», 6 novembre 2023.
Pause Giant AI Experiments: An Open Letter, Future of Life Institute, 22 marzo 2023.
B. Perrigo, *OpenAI Could Quit Europe Over New AI Rules, CEO Sam Altman Warns*, in «Time», 25 maggio 2023.
S. Piantadosi (@spiantado): «Sì, ChatGPT è straordinario e lascia a bocca aperta. No, @OpenAI non ha neanche lontanamente affrontato la questione dei bias…», Twitter, 4 dicembre 2022, 10.55, https://twitter.com/spiantado/status/1599462375887114240?lang=en.
K. Piper, *Sam Bankman-Fried Tries to Explain Himself*, Vox, 16 novembre 2022.
Rishi Sunak & Elon Musk: Talk AI, Tech & the Future, Rishi Sunak (canale YouTube), 3 novembre 2023.
Romney Leads Senate Hearing on Addressing Potential Threats Posed by AI, Quantum Computing, and Other Emerging Technology, www.romney.senate.gov, 19 settembre 2023.
K. Roose, *Inside the White-Hot Center of AI Doomerism*, in «The New York Times», 11 luglio 2023.
Sam Altman: 'I Trust Answers Generated by ChatGPT Least than Anybody Else on Earth', Business Today (canale YouTube), 8 giugno 2023.
P. Singer, *The Life You Can Save*, Random House, New York 2010 (tr. it., *La cosa migliore che tu puoi fare. Cos'è l'altruismo efficace*, Sonda, Casale Monferrato 2016).
Statement on AI Risk, Center for AI Safety, maggio 2023.
C. Vallance, *Artificial Intelligence Could Lead to Extinction, Experts Warn*, BBC News, 30 maggio 2023.
J. Vincent, *OpenAI Sued for Defamation after ChatGPT Fabricates Legal Accusations against Radio Host*, The Verge, 9 giugno 2023.
A. Weprin, *Jeffrey Katzenberg: AI Will Drastically Cut Number of Workers It Takes to Make Animated Movies*, in «Hollywood Reporter», 9 novembre 2023.
E. Yudkowsky, *Pausing AI Developments Isn't Enough. We Need to Shut It All Down*,

in «Time», 29 marzo 2023.

15. Scacco matto

The Capabilities of Multimodal AI | Gemini Demo, Google (canale YouTube), 6 dicembre 2023.
J. Dastin, K. Hu, P. Dave, *Exclusive: ChatGPT Owner OpenAI Projects $1 Billion in Revenue by 2024*, Reuters, 15 dicembre 2022.
M. Gurman, *Apple's iPhone Design Chief Enlisted by Jony Ive, Sam Altman to Work on AI Devices*, Bloomberg, 26 dicembre 2023.
K. Hagey, D. Seetharaman, B. Jin, *Behind the Scenes of Sam Altman's Showdown at OpenAI*, in «The Wall Street Journal», 22 novembre 2023.
M. Hawkins, E. Ludlow, G. Tan, D. Bass, *OpenAI's Sam Altman Seeks US Blessing to Raise Billions for AI Chips*, Bloomberg, 16 febbraio 2024.
A. Heath, *Mark Zuckerberg's New Goal Is Creating Artificial General Intelligence*, The Verge, 18 gennaio 2024.
A. Imbrie, O. Daniels, H. Toner, *Decoding Intentions: Artificial Intelligence and Costly Signals*, Center for Security and Emerging Technology, ottobre 2023.
C. Metz, T. Mickle, M. Isaac, *Before Altman's Ouster, OpenAI's Board Was Divided and Feuding*, in «The New York Times», 21 novembre 2023.
K. Roose, *Inside the White-Hot Center of AI Doomerism*, in «The New York Times», 11 luglio 2023.
M. Sigalos, R. Browne, *OpenAI's Sam Altman Says Human-level AI Is Coming but Will Change World Much Less Than We Think*, CNBC, 16 gennaio 2024.
J. Victor, A. Efrati, *OpenAI Made an AI Breakthrough before Altman Firing, Stoking Excitement and Concern*, in «The Information», 22 novembre 2023.
B. Walker, *Inside OpenAI's Shock Firing of Sam Altman*, Bloomberg, 20 novembre 2023.
M. Zuckerberg, *Some Updates on Our AI Efforts*, video pubblicato il 18 gennaio 2024 su Facebook, https://www.facebook.com/zuck/posts/pfbid02UhntmXwNBLiV8EZHK71gAQm

16. All'ombra dei monopoli

R. Bommasani, K. Klyman, S. Longpre, S. Kapoor, N. Maslej, B. Xiong, D. Zhang, P. Liang, *The Foundation Model Transparency Index*, Stanford Center for Research on Foundation Models (CRFM) e Stanford Institute for Human-Centered Artificial Intelligence (HAI), 18 ottobre 2023.
M. Cheng, *AI Girlfriend Bots Are Already Flooding OpenAI's GPT Store*, in «Quartz», 11 gennaio 2024.

M. Cheng, *A Startup Founded by Former Google Employees Claims that Users Spend Two Hours a Day with Its AI Chatbots*, in «Quartz», 12 ottobre 2023.
A. Holmes, *Microsoft CFO Says OpenAI and Other AI Products Will Add $10 Billion to Revenue*, in «The Information», giugno 2023.
Introducing the GPT Store, www.openai.com, 10 gennaio 2024.
K. Leswing, *Nvidia's AI Chips Are Selling for More than $40,000 on eBay*, CNBC, 14 aprile 2023.
The Long-Term Benefit Trust, www.anthropic.com/news/the-long-term-benefit-trust, 19 settembre 2023.
A. Schiffmann (@AviSchiffmann): «Ho appena realizzato l'AI indossabile più personale al mondo! Con Tab, potrete parlare di qualsiasi cosa. Da oggi, i nostri computer sono anche i nostri partner creativi!», demo di Tab, Twitter, 1° ottobre 2023, 5.12, https://twitter.com/AviSchiffmann/status/1708439854005321954?lang=en.
A. Vance, *Elon Musk's Brain Implant Startup Is Ready to Start Surgery*, in «Bloomberg Businessweek», 7 novembre 2023.

La vita di un libro non finisce con l’ultima pagina

C’è un lettore che sta aspettando il libro giusto da leggere: se questo ti è piaciuto, fai scattare la scintilla con un consiglio o con una recensione.

Per restare sempre aggiornato sulle novità del Gruppo editoriale Mauri Spagnol visita il sito e cerca sui social.

IL LIBRAIO